University of Pittsburgh

Evaluation Approaches for Political Party Assistance: Methodologies and Tools

Scott Morgenstern
University of Pittsburgh

Steve Finkel
University of Pittsburgh

Andrew Green
DGMetrics, Inc. and Georgetown University

Jeremy Horowitz
Dartmouth College

With special guidance from Barry Ames and Lou Picard and support from Miguel Carreras & Reynaldo Rojo

University of Pittsburgh

October 28, 2011

This study is made possible by the generous support of the American people through the United States Agency for International Development (USAID). The contents are the responsibility of the University of Pittsburgh and do not necessarily reflect the views of USAID or the United States Government.

Democracy and Governance-Elections and Political Processes (DG-EPP) not only for their funding, but also for their extensive feedback on every aspect of the project

Table of Contents

Evaluation Approaches for Political Party Assistance: Methodologies and Tools

Executive Summary

This document provides guidance on conducting evaluations of USAID's Political Party Assistance (PPA) programs. Obtaining data can be difficult in PPA activities, but there are alternatives for the adoption of rigorous approaches that are well suited for contributing to the stock of knowledge about "what works" in assistance for political parties and party systems. Many of the development concepts in this text reflect the companion piece, "Democracy, Parties, and Party Systems: A Conceptual Framework for USAID Assistance Programs."

Here we identify alternative methods and offer practical advice on how to use them in the context of USAID's political party assistance efforts. There is no one single evaluation type that will satisfy all of USAID's monitoring and evaluation needs in this sector. Rather, we believe that USAID ought to employ a variety of complementary evaluation approaches that will help USAID and its partners better understand when and how party assistance produces the intended results.

"Impact evaluations" (in which comparisons are made for development outcomes for individuals or other units that are the beneficiaries of some intervention and an appropriate and rigorously-defined control group) are the most credible approach for identifying the effects of specific programs. These types of evaluations are suitable for some programs, but we also provide guidance on how to conduct "performance" evaluations for projects and programs that are less amenable to these techniques.

We have two primary messages. Whether conducting randomized trials or performance evaluations, high-quality evaluations require comparison groups. They also require significant planning, beginning, when possible at the programming stage.

Methodological Overview

There are three distinct aspects of USAID's overall monitoring and evaluations. *Project monitoring* focuses on the implementation process and is designed to ensure that the recipients of USAID contracts and grants carry out the work they have agreed to perform. *Project Evaluations* are used to determine whether USAID-funded projects and programs have their intended effects and to learn about how best to achieve results. Finally, *country assessments* are used to examine conditions within a particular country regarding democracy and governance issues.

For monitoring or evaluations, we differentiate among four levels of indicators to measure performance. *Output* indicators count the specific activities carried out by a project, such as the number of training sessions held or the number of technical assistance workshops conducted. *Outcome* indicators track the direct results that are expected to come from a particular project, such as an improvement in parties' ability to develop platforms that reflect voters' concerns. *Impact* indicators measure the larger effects of projects at the

sector level, such as behavior changes. Finally, *higher-level impacts* refer to indicators that measure the results of behavior or policy changes.

Evaluation Challenges

The "fundamental problem of causal inference" states that the true effect of any intervention is an intrinsically unobservable quantity. Thus, evaluations must be designed so that *observable* quantities can serve as proxies for the unobservable "what if" or counterfactual untreated state for the units that receive treatment. This requires disentangling the effects of USAID-funded activities from the many other "confounding" factors that also influence aspects of party and party system development that are of interest in the study. Identifying the causal effects of USAID-funded (or other) interventions requires collecting trend data (baselines and endlines) and the use of control groups. In some cases, however, baseline data will be absent and it will not be possible to reconstruct them.

A second challenge for evaluations is the potential the USAID programs to have unintended consequences. Generally speaking, party development implies advancement along two axes: representation and governability. Advances along one axis may work against improvements along the other. Thus, evaluators need to determine whether USAID-funded projects produce both the intended effects, and whether they lead to unintended effects.

Next, in some cases, a project's effects will not be visible in the short-term. In principle, evaluations should be appropriately timed in order to capture the effects of a project. In practice, however, it may be difficult or impossible for USAID to evaluate a project many years after its conclusion.

There are three other challenges. First, the magnitude of effects of USAID projects may vary in different contexts. The next is stems from the assumption that treatment effects are "stable," meaning that a treatment *d* given to one unit is the same as treatment *d* given to another unit, and that the effect of a treatment *d* on a given unit does not depend on how treatments may be assigned to any other unit. Finally, evaluators need to be sensitive to the size of the intervention. In some cases, projects may fail to achieve their desired effects not because the project is poorly conceived but because its scope or reach is too small. Thus, evaluators will need to determine whether USAID funding levels are sufficiently large to achieve the stated goals.

Evaluation Approaches

We describe two basic evaluation approaches, with several subtypes for single-country studies. There are several types of impact evaluations, experimental, longitudinal, and cross-sectional. When these types are not feasible, we prescribe a two-pronged approach to "performance" evaluations, designed to maximize their value in light of the inferential challenges they face. Regardless of the type of evaluation, we emphasize the importance of identifying a comparison group to improve the quality of the data.

Evaluations with Random Assignment of Treatment

The evaluation of PPA programs may involve "Randomized Control Trials" (RCT), where some aspect of a USAID program, or "treatment," would be randomly assigned across

individuals, groups, geographic areas, or other units. That is, some units would receive the treatment and other units would not, with the assignment mechanism being a random chance. The resultant treatment and control groups will be statistically equal (given large enough samples) in every observable and unobservable way except that the treatment group received the program's intervention and the control group did not. While the inability to randomly assign treatment does not preclude meaningful evaluation, the strongest and most credible inferences are made when comparing treatment and control group within the context of a randomized controlled trial

Two types of activities in particular are well suited for randomization – those that involve a large number of geographical units (villages, districts, constituencies, etc.) within a particular country, and those that involve a large number of individual participants. Examples include "oversubscription" of individuals in a training program and support of party activities in some of a country's regions.

Randomized evaluations do face faces difficulties. One is the potential for "spill-over effects," or interference between the control group and the treatment group. Additionally, the behavior of treated or control individuals may change because treatment was applied or withheld.

Evaluations with Non-Random Assignment of Treatment

Sometimes it is not feasible to evaluate a program through randomized treatment assignment. One alternative methodology employs over-time or *longitudinal* comparisons of the treatment and control groups. These studies use non-randomly assigned treatment and control units for a single point in time. The potential for selection bias is severe, but two strategies may be useful to overcome these situations. One involves "matching" treatment and control units on observed factors that distinguish the treatment and control groups, calculating the difference in some outcome (impact) within the matched strata, and aggregating these differences into an overall "treatment effect." The second involves adding information through "instrumental variables" into the analysis.

Performance Evaluation

When evaluations are commissioned "after the fact," it may not be possible to construct valid control groups and in some cases it will be impossible to assemble time series data that stretches back to the beginning of the project. For these cases, we suggest –to the extent possible- that evaluation teams construct comparative groups by collecting information about the pre-treatment period or about individuals or regions that did not participate in the programs. Because these types of studies will generally lack carefully constructed control groups, confounding factors are a special concern. These types of evaluations, then, must carefully consider contextual and other factors as potential influences on the parties and party systems.

We suggest that these types of evaluations include two complementary components. The first one starts with program activities and seeks specific evidence to determine direct results from the activities. The second component works from the perspective of a role of parties and party systems in a developed democracy (accountability, representation, and

participation, governability and good governance, etc.) and seeks to determine whether the country has made progress toward these aspects of development.

To facilitate and systematize these studies, we provide several working tables and appendices. For the "by-activity" task, the tables recommend that evaluation teams enumerate the activities and identify indicators for expected outputs, outcomes, and impacts. For the by-development characteristic work, the team should list the development characteristics, note whether they were expected targets of the program, and the level of effort expended in addressing the issue. It would then detail the activities associated with each development characteristic and data needed for evaluation (including control groups).

Summary and Recommendations

While political party aid creates some special evaluation challenges, careful attention to methodology can greatly improve the quality of information that impact and performance evaluations provides. The report concludes with five main points:

- USAID should define outcome, impact, and higher-level impact variables in advance of program implementation.
- USAID must place greater emphasis on collecting high quality data that can be used to measure key trends in the intended outcome and impact variables.
- Evaluations must include observations on "control groups" that were unaffected by PPA interventions, and not simply rely on observation of units that were "treated."
- USAID should design Political Party Assistance programs while taking into account evaluation considerations to a greater extent. This includes constructing control groups using random assignment of treatment by taking advantage of different randomization strategies, collecting pre- and post-intervention data for treated and untreated units when randomization is not possible
- Evaluations must be sensitive to the needs for balancing methodological validity with judging higher-level impacts.

Introduction

This document provides guidance for USAID, its implementing partners, and teams commissioned to conduct rigorous evaluations of Political Party Assistance (PPA) programs. These programs are complex, and careful monitoring and evaluation is necessary in order to provide appropriate evidence for improving results based program design. Monitoring and evaluation, thus, aids the success of development programs.

Evaluators seeking to identify the effects of party assistance activities face a number of challenges – some common to the broader Democracy and Governance field, others unique to the PPA sector. Obtaining data can be difficult in PPA activities, particularly when evaluators must rely on the beneficiaries of assistance (political parties) to provide information. In practice baseline data is often missing and difficult to reconstruct retrospectively. And in settings that lack control groups and/or baseline data, it can be difficult to link USAID programs to trends in party system development over time. However, in spite of these manifold challenges, opportunities exist for the adoption of more rigorous approaches that are well suited for contributing to the stock of knowledge about "what works" in assistance to political parties and party systems.

Many of the development concepts in this document reflect our companion piece, "Democracy, Parties, and Party Systems: A Conceptual Framework for USAID Assistance Programs" (hereafter *Conceptual Framework*). In developing the recommendations outlined below, we draw particularly on the 2008 report by the National Academy of Science, *Improving Democracy Assistance: Building Knowledge through Evaluation and Research* (hereafter NAS 2008), and on the USAID document, *Evaluation Policy*, produced in January 2011 by the Bureau of Policy, Planning, and Learning. In line with those reports, we offer an array of evaluation approaches. We identify the strengths and weaknesses of alternative methods and offer practical advice on how to use each type in the context of USAID's political party assistance efforts. There is no one single evaluation type that will satisfy all of USAID's monitoring and evaluation needs in this sector. Rather, we believe that USAID ought to employ a variety of complementary evaluation approaches that will help USAID and its partners better understand when and how party assistance produces the intended results.

The NAS 2008 Report and the 2011 USAID Evaluation Policy identify "impact evaluations" (in which comparisons are made for development outcomes for individuals or other units that are the beneficiaries of some intervention and an appropriate and rigorously-defined control group) as the most credible approach for identifying the effects of specific programs. Moreover, it is widely recognized that impact evaluations that are characterized by *random assignment* of units to either the "treatment" or "control" group "provide the strongest evidence of a relationship between the intervention under study and the outcome measured"(*USAID Evaluation Policy,* p.2). As such, we will discuss evaluations with randomized treatments in some detail below, and provide guidance on when and how such techniques can best be utilized to evaluate PPA activities. It will be seen, for example, that randomized evaluations are most suitable for programs that involve a large number of geographical units (villages, districts, constituencies, etc.) within a particular country, and those that include large numbers of participants, for example, in training programs, as

those kinds of programs present relatively few obstacles to successful randomization of treatment and control units.

Some PPA activities, however, are not as easily amenable to randomized evaluations. For example, it is generally unethical and against USAID regulations to exclude some parties from programs for the sake of randomization. It would also be unethical to exclude particular leaders, such as in a program that emphasized private meetings with leaders to encourage coalition building, responsible behavior of opposition parties, or support for internal reforms. In these types of cases it would be difficult (but not always impossible as we discuss below) to construct an appropriate control groups against which to compare the outcomes of the "treated" or beneficiary group. Such goals, further, do not lend themselves to developing clear indicators that could show the degree of program success. Nevertheless, a number of potentially effective designs exist for conducting non-randomized impact evaluations, most notably "difference-in-difference" or other kinds of longitudinal (over-time) designs whereby treatment and control groups are observed over time in order to parse out the effects of a given intervention. Along these lines, comparisons among different regions of a county, some of which may have been "treated" more intensively with PPA activities than others, or indeed comparisons between different countries with different levels of PPA treatment over time can also be fruitful in assessing effects. These latter comparisons, in fact, may be most beneficial for demonstrating the "higher level" impacts of PPA programs and the effects of overall or aggregate levels of PPA activities.

Beyond these difficulties, when evaluations are commissioned after the fact—as is common— to try to determine whether a PPA activity (or bundle of activities) has had the intended effects, the difficulties are magnified. For single-country performance evaluations, for example, USAID has typically relied on the "three person, three weeks" model, whereby a small number of external consultants (often three) are contracted to evaluate a recently completed program by spending a relatively short amount of time (often three weeks) conducting in-country interviews and gathering data. This model is poorly suited for formal impact evaluations, because the information that can be obtained through interviews is necessarily anecdotal and incomplete, and the amount of time spent in country is often inadequate for collecting detailed data on key trends. The study team will also have difficulty in generating information about baselines or the comparison groups that are a necessary piece of impact evaluations. Nevertheless, there is room for performance methodologies (using both qualitative and quantitative information) to be part of PPA's evaluation toolkit, and we therefore provide guidance on ways to improve this type of analysis. Again emphasizing the importance of comparative data, we outline a two-pronged approach that is designed both to identify evidence of effects of the specific activities funded by USAID while also tracking progress toward broader program development goals over time. In this way, the evaluation provides a comprehensive picture both of direct effects at the most proximate level while also offering evidence about higher-level impacts to which USAID-funded activities may be linked.

Regardless of the type, it is important to underscore the scope of these evaluations and the role of evaluations in improving development programs. Our purpose is to describe

methodologies for analyzing of the effects of programming, rather than emphasizing explanations for the why a program had (or failed to have) a particular effect. That type of information is clearly important to future programming; thus while the subject is somewhat beyond our scope, two points are in order. First, in order to pursue that type of information, analysts would have to first substantiate the effects by conducting the types of evaluations we suggest here. Second, while the examples in this document focus on measuring overall effects rather than explanations, the methods we suggest here can be applicable to explanatory analyses as well. For example, a randomized study could show which types of program designs (e.g. single or multi-party trainings) were most effective or whether the effect of programs was affected by the electoral calendar, regional conditions, program cost, or other factors.

A second issue is also beyond our scope: the relationship between the evaluation team, USAID missions, and program implementers. While those responsible for implementing the development programs will be interested in learning about the success of their programs, evaluations may raise some apprehensions. Our caution, in short, is that the evaluation team will have to consider these kinds of sensitivities.

We have two primary messages in this document. First, whether conducting randomized trials or performance evaluations, high-quality evaluations require comparison groups. Second, they also require significant planning, beginning, when possible at the programming stage. At that stage USAID and its implementing partners can determine expected outcomes, devise indicators, and consider control groups. As part of this process, they may also decide what programs will not be the object of later evaluation. Post-hoc evaluations will also improve with significant strategic planning. As we describe later, the quality of evidence that these evaluations can provide improves greatly when the evaluation team defines comparative groups and specifies indicators that show change.

The document is structured as follows. Part 1 provides a methodological overview and framework for analysis. It begins by defining key terms. It then outlines a number of core challenges that evaluators are likely to face in studying the effectiveness of PPA programs. Part 2 describes the menu of evaluation strategies alluded to above, ranging from randomized to non-randomized impact evaluations to more qualitative performance studies. This part is designed to serve as a toolkit for evaluators, providing guidance on when and how evaluation methodologies of different kinds can be used and to provide concrete examples to illustrate the potential of the alternative approaches. Part 3 concludes with recommendations for future evaluations, and with recommendations for the design and implantation of future programs so as to better facilitate the evaluation of USAID PPA activities.

Part 1. Methodological Overview

1.1 Definitions

It is important at the outset to define several key concepts that are used throughout this document. First it is important to differentiate between three distinct aspects of USAID's overall monitoring and evaluations efforts: project monitoring, evaluations, and country

assessments (USAID *Evaluation Policy, 2011;* NAS 2008, pp. 60-62). *Project or performance monitoring* focuses on the implementation process and is designed to ensure that the recipients of USAID contracts and grants carry out the work they have agreed to perform. Monitoring is a routine, on-going function that is currently carried out through USAID's performance monitoring system. *Evaluations* are used to determine whether USAID-funded projects and programs have their intended effects and to learn about how best to achieve results. Evaluations are conducted to provide USAID with "basis for judgments, to improve effectiveness, and/or inform decisions about current and future programming" (*USAID Evaluation Policy,* p. 2). Increasingly USAID is requiring implementers to develop and integrate evaluation procedures into their project plans from the outset, based on the recognition that rigorous evaluations must be built in at the project design phase. Finally, *country assessments* are used to examine conditions within a particular country regarding democracy and governance issues. Assessments are used by USAID to design programs that are tailored to a particular country context. USAID conducts both broad country assessments that look at an array of indicators (these are known as Democracy and Governance assessments), as well as sub-sector assessments that focus, or example, on civil society, political parties, or the judicial system.

It is also important to differentiate between the main types of indicators used in USAID's monitoring and evaluations efforts – outputs, outcomes, and impacts.[1] As shown in Table 1, *Output* indicators are used to count the specific activities carried out by a project, such as the number of training sessions held, the number of technical assistance workshops conducted, or the number of people who participated in a study tour to a foreign country. *Outcome* indicators track the direct results that are expected to come from a particular project, such as an improvement in parties' ability to develop platforms that reflect voters' concerns, the frequency with which candidates visit their home districts, or the number of pro-reform laws passed. *Impact* indicators measure the larger effects of projects at the sector level. Typically impact indicators measure behavioral changes, as when party leaders select candidates in a more transparent fashion, raise funds through legal means rather than illicit processes, or engage more frequently in negotiations across partisan lines. Finally, we define *higher-level impacts* to refer to indicators that measure the results of behavior or policy changes.

To illustrate these distinctions in concrete terms, we offer three examples of party assistance projects. A common program activity, displayed here as Example 1, is to help guide dialogues in support of NGOs or the parties themselves in their efforts to pursue consensus about needed reforms. A problem that many countries face is political party system fragmentation. To support a process of election law reform, NDI or IRI might support a series of workshops that bring together party leaders, civil society groups, and outside experts to discuss the effect of electoral laws on the party system. In this case, we define the output as the number of workshops held. There are two outcome indicators – whether there was an increased understanding of electoral law issues, and whether consensus about the need for reform and the direction of reform was achieved. The impact

[1] The OECD's Development Assistance Committee (DAC) and the *USAID Evaluation Policy* use related but distinct definition of these terms.

indicators would include whether a bill were drafted and debated in the legislature and perhaps whether the reforms were actually enacted (i.e., a new law was passed). Higher-level impact indicators measure changes in party behavior – whether the new law had its desired effect, in this case (a legal framework that is conducive to a more stable and consolidated political party system) in part by reducing the number of parties. Note that in this case the anticipated impact of reducing the number of parties was a goal of the parties, not of the donor. It may have been apparent at the outset, however, that the reason for supporting the conferences was that the parties were concerned with fragmentation. Also note that there is not always a clearly demarcated line between outcomes, impacts, and higher-level impacts. Passage of the law, for example, could be construed as a higher-level impact because enactment requires much more than a simple increase in knowledge.

Example 2 is a project that seeks to increase the number of female legislators in a country's parliament. The project offers a series of training sessions that provide female candidates with technical assistance on how to conduct more effective campaigns. Output indicators [2]measure both the number of training sessions held and the number of women who participated. At the outcome level, we are interested in whether learning occurred and whether participants were able to carry out new activities based on their increased knowledge. For example, if the training program offered guidance on how to collect information from voters on their priorities through opinion polls, one outcome of interest would be whether the participants did in fact develop and implement surveys. Impact indicators for this project would measure whether participants were in fact able to conduct better campaigns as a result of the training. This could be measured by examining whether the participants were better able to raise funds, develop platforms that reflect voter concerns, and give better stump speeches. The higher-level impact indicators measure whether more women chose to run for office as a result of the project and whether more female candidates succeeded in getting elected.

As a third example, consider a conference or a series of training programs to teach parties about strategic planning (perhaps as a means to emphasize incorporating policy goals into the party's platform or to plan outreach campaigns). In this example the output variable would again be the number of participants or parties trained. The outcome could be the outlining of a strategic plan during the conference or training. There would be an impact if the participants then developed their strategic plan, after the training had ended. Gauging the higher-level impact would require answering questions about the parties' behavioral changes and, more importantly, whether the new plan helped them attract more members, raise more funds, or add policy content to their campaigns.

[2] Note that we follow NAS and other USAID reports and refer to "impact evaluations" but also use the word "impacts" as one of three levels of indicators, as defined here.

Table 1: Indicator Types

	Output	Outcome	Impact	Higher level Impacts
Definition	Indicator focused on counting activities	Indicator focused on direct, proximate effects of project	Indicator focused on behavior change effects from project	Indicator focused on results of behavior or policy changes
Example 1: Workshops on electoral law reform	1. Number of workshops held	1. Electoral laws understood 2. Consensus built	1. New law debated in legislature 2. Law adapted	1. Law has intended effects (Number of parties decreases)
Example 2: Training for female candidates	1. Number of trainings held 2. Number of participants	1. Learning 2. Questionnaire developed and implemented	1. Women run better campaigns – e.g., raise more funds, develop platforms that better reflect voter concerns, give better campaign speeches	1. More women run for office 2. More women get elected
Example 3: Training to develop parties' strategic plans	1. Number of trainings held 2. Number of participants	1. Sample plan developed during training	1. Full and effective plan developed and adopted as party policy after training ended	1. Parties alter behavior by changing outreach, fundraising, or other practices 2. Membership or fundraising improves; More policy is incorporated into campaign

1.2 Evaluation Challenges

Evaluations in the political party assistance sector face a number of important challenges. Some of these challenges are common to all impact evaluations, in that there are a series of inherent difficulties in ascribing causal effects to any social or political intervention. Some other challenges, though, are more specifically relevant to PPA programs, given difficulties in designing evaluations to assess programs' effects in both the short and long terms, difficulties in data collection on party-relevant outcome indicators, and given uncertainties regarding how effects may differ across different political and social contexts, and across

programs of different sizes. The goal of this section is to alert evaluators (and those who commission evaluations) to these fundamental challenges, and to set the stage for discussing in the next section the methodologies that will enable evaluators to overcome them to as great an extent as possible.

- *"The Fundamental Problem of Causal Inference"*

Evaluations of the impact of social or policy interventions, whether they are programs designed to alleviate poverty, improve educational outcomes, or to strengthen political parties in developing democracies, face a number of significant challenges. Foremost among them is the so-called "fundamental problem of causal inference" (Holland 1986), which states that the true effect of any intervention is an intrinsically unobservable quantity. That is, what we seek to estimate in impact evaluation is the difference between some clearly-defined outcome (or impact) for a unit that is "treated" by some intervention and what the same clearly-defined outcome *for the same unit* would have been, were it not treated by the intervention in question. Because it is only possible at any given time to observe a unit in *either* a "treated" or "untreated" state, the true effect of an intervention for any unit, and hence the aggregated average treatment effect across all units, is essentially unknown.

How, then, can impact evaluation proceed? Given the "fundamental problem," evaluations must be designed so that *observable* quantities can serve as proxies for the unobservable "what if" or counterfactual untreated state for the units that are treated by a particular intervention. These quantities are measurable outcomes (or impacts) from a collection of *observed* untreated units or a "control group", and the effect of an intervention is then estimated by comparing in some statistical fashion the average outcomes (or impacts) for the treated units to the average outcomes (or impacts) for the control units.

There are several important implications from this basic exposition. First, meaningful estimates of a program's impact depend on the existence (or construction) of some kind of control group that can provide readings of what the treated units would have looked like in the absence of treatment. Without such observations, the intrinsically unobservable causal effect of an intervention – that is, how different at a given point in time the treated units would have looked in the absence of treatment –remains hopelessly unobservable. This means, for example, that one cannot estimate the impacts of a program designed to strengthen campaigns run by women candidates by examining only the campaigns of candidates who participated in the program; their campaigns may have been strong without having had the benefit of the program's training[3]. Similarly, one cannot estimate the impacts of programs designed to promote party decentralization by examining the local-level organizations of parties only in areas where the program operated; again, the decentralization that those parties exhibited may well have been high in the absence of

[3] As will be seen below, it is also not sufficient to estimate program impact by examining pre and post treatment outcomes only for the treatment group. That is, what the treatment group looked like before treatment is *not* a sufficient counterfactual condition for valid impact estimation, since whatever changes the treatment group experienced pre and post treatment may have also been observed over that same time period among units that were not treated. Again, observations on untreated units are essential for making appropriate estimates of program impact.

treatment as well. Thus credible estimation of program effects depends on the existence of *some* kind of control group that can serve as the counterfactual "untreated state" for the treated units. In the examples above, control groups could be women candidates who did not participate in the campaign training activities, and levels of decentralization for parties in regions where the program did not operate.

Second, comparisons of treatment and control groups are only valid estimates of the true effect of an intervention to the extent that the control group is comparable to the treatment group – that is, statistically in all respects aside from treatment status. It is for this reason that randomization of treatment is the preferred method for estimating causal effects, since it ensures that, with a large enough sample of units, the treatment and control groups will be equated on *all* potentially confounding factors,[4] those that the researcher can identify and measure in advance, and even those which may be unknown or unmeasurable as well. Consider the example of the women's campaign strengthening program. Women candidates who volunteer (or "self-select") into the training program may be quite a bit different in terms of skills, networks, resources, and other hard to measure characteristics of women candidates who did not participate in the program, and hence a simple comparison of "average campaign quality" between these groups could be a seriously misleading measure of the program's impact. Randomization, in theory, solves this problem, since a large enough control group of randomly selected women candidates will be *identical* to the randomly selected "treated" women candidates on all measured and unmeasured characteristics aside from treatment status.

When randomization of treatment is not possible, the *a priori* equality of treatment and control groups is not guaranteed, and valid comparisons between the groups are only possible to the extent that certain assumptions can be justified within the context of a given evaluation design.

For example, assume, in the case of self-selected women candidates for a campaign training program, that an evaluator believes that the sole differences between women candidates who volunteered for the program and those who did not were in their levels of prior resources such as education and income. She then estimates treatment effects by "matching" treated and untreated candidates on these characteristics, and comparing campaign quality within all of the matched strata. This comparison, however, will be valid only to the extent that the treated and untreated women do not differ on factors aside from resources that may also relate to campaign outcomes. It seems plausible that self-selection into such training programs would be more prevalent among women who had higher levels of campaign-relevant (and potentially unobservable) qualities such as intrinsic motivation, verbal skills, and sociability, and to this extent, inferences about the impact of the program would be biased. Again, this points to importance of careful design and *a priori* thinking about how appropriate control groups can be constructed for a given evaluation.

- *Identifying Program Effects on Impacts and Higher-level Impacts*

[4] In the language of causal inference, social scientists discuss use the term "confounding factors" to imply forces that affect outcomes other than the treatment. A new electoral law, international trends, wars, and new sources of funding are examples of potentially confounding factors affecting party development.

The foregoing discussion illustrates the intrinsic difficulties in estimating the effects of PPA activities even on the immediate, proximate *outcomes* ---e.g., changes in workshop participant attitudes or knowledge, or increased consensus on reform needs among program participants from different parties -- that the programs are designed to achieve. These challenges are magnified to an even greater extent when evaluators attempt to assess program effects on what we term *impacts* and *higher-level impacts*, e.g., actual changes in electoral laws, or changes in the number of women candidates in a particular country. Assessing program effects on these kinds of impacts involves the additional challenge of disentangling the effects of USAID-funded activities from the many other factors that also influence the aspects of party and party system development that are of interest in the study. In addition to USAID's activities, other international organizations also fund similar programs, and domestic political forces have great influence. New laws, rising political movements, shifting political ideologies, or new ideas imported from neighboring countries all lead parties to adjust their strategies, behaviors, and institutions.

To see this challenge in concrete terms, we return to the party law reform example introduced above. To reiterate, the program sought to support discussions about reforms due to high fragmentation in the party system. Even if the program were successful in changing an outcome such as generating consensus among party participants, multiple factors could upset the program achieving its intended *impact*, the translation of that consensus into a new law. Key dissenting parliamentarians, for example, could derail the reform, or societal groups might demonstrate in opposition to the proposed reform if they saw threats to their representatives. The problem becomes worse as we move to the *higher-level impacts*. If a new law were passed with provisions that, for example, raise the threshold for parties to gain seats in the parliament, the expected effect would be a reduction in the number of small parties competing for votes over time. To test for effects, evaluators might simply track the effective number of parties (a weighted measure of the number of parties) to see if the number did in fact decrease after the introduction of the law. Attributing the change to the programs, however, is problematic, because of the multiple factors that affect the number of parties over time. Such factors could be influential enough to prevent change even in the face of a solid program. In the language of causal inference, social scientists discuss these concerns in terms of *confounding factors*. In the present example, there are many possible confounds that might be at work. Three possible confounding factors are shown in Figure 1. Even if the new law did encourage some parties to merge (in order to surpass a new threshold, for example), large parties might split or new parties might form. These changes could occur in response to other legal changes, such as reform to campaign finance laws, but they may also occur in response to a new salient political issue that leads to splits in old parties or to new party formation. Due to these types of confounding factors, the number of parties might stay the same or actually increase following the introduction of the party reform initiative. In this case, the program may have been successful in promoting a new law, but the desired end result was undermined by other factors. This is also related to the concern with the "level of effort;" it may be unrealistic to expect large-scale changes from the relatively small-scale USAID programs. As such, evaluations might mistakenly conclude that a program was unsuccessful, where in fact the program helped push in the desired direction, but was

overwhelmed by strong countervailing pressures. In sum, a focus on impacts or higher-level impacts may lead to mistaken inferences about the role of USAID (or other) programs.

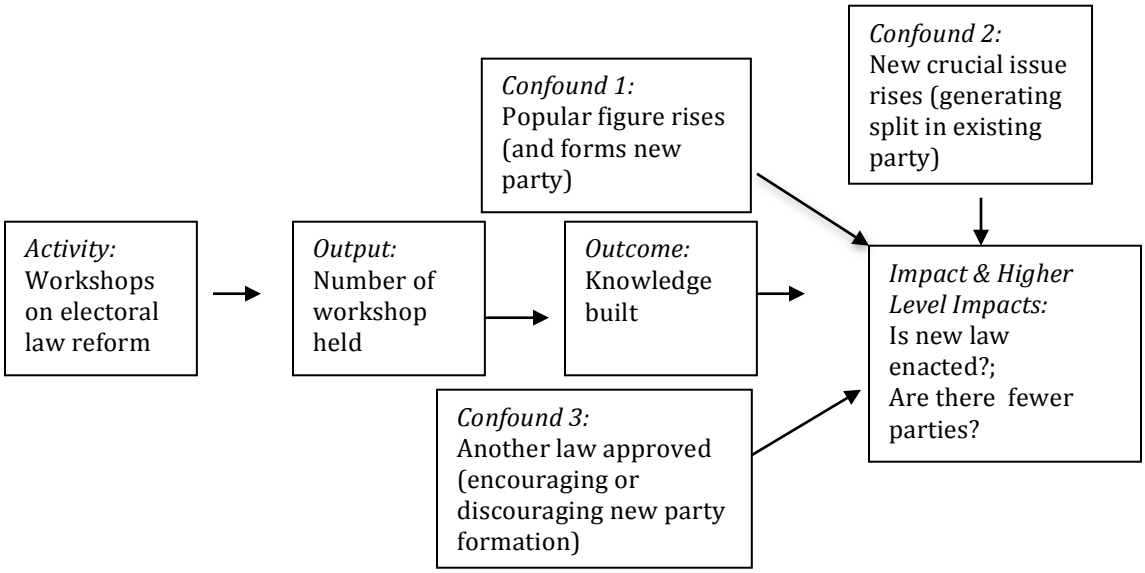

Figure 1: Electoral Law Reform and Confounding Factors

The foregoing discussion has several implications for the practical evaluation of USAID PPA activities. First, it should be "easier," relatively speaking, to assess effects on project *outcomes* compared to effects on *impacts* and *higher-level* impacts. Second, it may be the case that a project has significant observable effects on outcomes, but its effects on *impacts* and *higher-level* impacts are not detectable, given strong countervailing conditions or confounding factors operating in a given region or country at a given point in time, and given the relative size of most USAID PPA activities.[5] To assess impacts at higher-levels, then, it may be necessary to expand the evaluations to include other regions or other countries which vary on their level of both PPA "treatments" and vary on important confounding factors so that they can serve as suitable counterfactual control groups. In cases where this strategy is not feasible, evaluators will have to deal with confounding factors by carefully considering the range of possible confounds that might be at work in a given setting and accounting for how each may affect impact variables. In such cases evaluators will generally be unable to determine with a great deal of certainty whether sector-level changes – be they positive or negative – can be attributed directly to USAID-funded projects rather than other factors. Evaluators may be limited to determining whether USAID programs appear to be working with or against these other confounding factors. Our discussion of ex-post (after the fact) evaluations below addresses these issues more fully.

[5] We note, however, that the reverse situation is highly unlikely to obtain. It is unlikely that a program has *no* significant effects on proximate outcomes but nevertheless influences impacts and higher-level impacts. This justifies a focus on determining outcome-effects of programs via rigorous impact evaluations, as effects at this level are very nearly a necessary condition for effects at higher levels. At the same time, finding effects on outcomes but not on higher-level concerns might suggest that project developers should reconsider their programming focus.

A final related question is whether the evaluations should focus on individual activities, on clusters of activities, or on the whole development program. In other words, while output and perhaps outcome indicators focus on the effects of individual activities, impacts and higher-level impacts reflect, to some extent, the success of a complete development program. In many (if not most) cases, programs include multiple inter-connected activities aimed at improving a single development deficit. For such cases, analyses must consider wider impacts to develop a complete evaluative picture.

- *Measurement / Data Collection Problems*

There are also a number of problems related to measurement of effects and data collection. Identifying the causal effects of USAID-funded (or other) interventions with a high degree of certainty requires the collection of trend data (baselines and endlines) and the use of control groups. In some cases, baseline data will be absent and it will be impossible to reconstruct baseline measures retrospectively. While USAID has increasingly required that baselines are established at the start of projects, in practice evaluators may be called upon to carry out evaluations where such data is not available. In some cases, evaluators may also be challenged by the privacy of political parties. Effective party support programs require trust between the party leadership and implementers, and public evaluations could reveal secrets that harm the relationship. At the same time, publicly available evaluations may be necessary for studies designed to reveal larger lessons about successful practices. Before undertaking the evaluation, then, USAID and implementers should agree on what types of information will be public.

- *Unintended Effects*

USAID programs may have unintended consequences. In broad terms, party development implies advancement along two axes – representation and governability (this point is developed more fully in the *Conceptual Framework*). Advances along one axis may work against improvements along the other. For example, a project that encourages the creation of more political parties may enhance representation of minority groups, but in doing so it may harm efficient decision making in the legislature. A project that encourages internal party democracy may improve accountability but may also create a system in which popular but unqualified leaders are chosen as candidates. Evaluations will face the challenge of determining whether USAID-funded projects produce both the intended effects, and whether they lead to unintended effects.

- *Long-Terms Effects*

In some cases, a project's effects will not be visible in the short-term. Consider an outreach and training program that seeks to develop emerging party leaders. It may take a decade or more before these young leaders work their way up through party ranks. In this case, the effects of the project may not be immediately visible. In principle, evaluations should be appropriately timed in order to capture the effects of a project. In practice, however, it may be difficult or impossible for USAID to evaluate a project many years after its conclusion.

- *Context Matters*

In seeking to measure the effectiveness of USAID projects, evaluators need to be aware that the magnitude of effects may vary in different contexts. As described in the *Conceptual Framework*, party assistance programs are likely to have different degrees of success in different political context. In newer democracies where institutions are weak and party system are still in formation, USAID projects are likely to have greater effects than in countries where institutions and party systems are more fully consolidated. At the same time, some less-consolidated democracies may be more resistant to outside influences. Evaluators need to expect these kinds of "heterogeneous treatment effects" (as they are termed in the evaluation literature), and implement appropriate designs and methodologies to detect them.

- *Variability in Treatments and the Problem of "Spillovers"*

An additional set of challenges stems from the assumption in classical evaluation theory that treatment effects are "stable," meaning that a treatment *d* given to one unit is the same as treatment *d* given to another unit, and that the effect of a treatment *d* on a given unit does not depend on how treatments may be assigned to any other unit[6]. In the case of PPA evaluations, it may be difficult to justify both aspects of the stability assumption. In the first place, programs are very likely to vary in how they are executed in practice, as different implementing partners will have different levels of resources, training and expertise, and different sites will be easier or harder locations in which programs may operate. To the extent that these factors are not known or taken into account by the researcher, the potential for erroneous inferences about "the" causal effect of a particular program will exist.

Second, it is unlikely that the effects of a given program will be completely independent of how treatment and control groups are assigned or allocated. For example, there is a possibility that individuals from a political party who are treated with a given intervention may nevertheless come into contact or influence individuals from that party who were not treated; these kinds of "spillover" effects or "interference" between the treatment and control groups mean that simple comparisons between (even randomly assigned) treated and untreated units may be insufficient to uncover a program's causal effects. Again, these possibilities need to be incorporated formally into a study's design in order to arrive at valid causal inferences, and the plausibility of the "stable" treatment effects assumption always needs to be justified in a given instance.

- *The Size of the Intervention*

As noted, evaluators also need to be sensitive to the size of the intervention. In some cases, projects may fail to achieve their desired effects not because the project is poorly conceived but because its scope or reach is too small. Thus, evaluators will need to consider the magnitude of particular interventions and determine whether USAID funding levels are sufficiently large to achieve the stated goals of the activities.

[6] This refers to SUTVA, or the "Stable Unit Treatment Value Assumption."

Part 2. Evaluation Approaches

This section provides an overview of basic options for evaluating the effects of party assistance programs, as summarized in Table 2. For single-country evaluations, we begin with those in which the party assistance project or some aspect of it is randomly assigned in order to create control and treatment groups. While this strategy can be used both within a single country and across many countries, we focus on the single-country approach because it has the most relevance for party assistance programs. Second, we describe evaluation methodologies that can be used when random assignment is not feasible. There we offer two strategies, one based on longitudinal (over-time) analyses and the other based on cross-sectional (comparisons among treated and untreated units). Finally, we discuss "performance" evaluations, which are common in USAID practice. These studies are commissioned after a program has taken place and they will therefore generally lack the benefits of having a clearly defined control group. We prescribe a set of procedures based on a two-pronged approach that is designed to maximize the value of such evaluations in light of the inferential challenges they face. As illustrated in the table, we emphasize throughout the importance of identifying a comparison group to improve the quality of the evaluation.

Table 2: Evaluation Types and Relevant Comparison Groups

Evaluation Types	Comparison Groups
Impact Evaluation Types	
Experimental	Randomized treated versus non-treated individuals, parties, or regions
Quasi-Experimental	
Longitudinal	Non-Randomized; compare pre- and post treatment, based on established baselines
Cross-sectional	Non-Randomized, "single shot" comparison of treated vs not-treated regions, individuals, or parties
Performance	
By Activity	Generally lacks pre-established comparison group, but may attempt to measure over-time or cross-regional change of treated individuals, parties, or regions
By Development Characteristic	Also lacks pre-established comparison group, but may try to establish indicators of change in development from pre-treatment period or differences between treated and non-treated areas

2.1 Evaluations with Random Assignment of Treatment

One option for the evaluation of PPA programs is to conduct what are referred to as "Randomized Control Trials" (RCT), where some aspect of a USAID program, or "treatment", would be randomly assigned across individuals, groups, geographic areas, or other units. That is, some units would receive the treatment and other units would not, with the assignment mechanism being a lottery or random chance. Following the logic of the "fundamental problem of casual inference" described above, the resultant treatment and control groups will be statistically equal (given large enough samples) in every observable and unobservable way except that the treatment group received the program's intervention and the control group did not. The control group's average value on a party-related outcome or impact indicator can then serve as a valid "counterfactual" to what the treatment groups' values would have been in the absence of treatment, and any difference on outcome or impact indicators between the treatment and control groups can be attributed to the intervention and not to confounding factors. While the inability to randomly assign treatment does not preclude meaningful evaluation, the strongest and most credible inferences are made when comparing treatment and control group within the context of a randomized controlled trial (NAS 2008, p. 5). Because of this, USAID and other donor agencies have begun in recent years to embrace integrating random assignment in program design as an important component of their overall monitoring and evaluations efforts (see, e.g., Moehler 2011).

While random assignment cannot answer all types of question pertaining to program evaluation nor serve as the only evaluation approach within the larger evaluation "toolkit," there are a number of types of PPA activities that are amenable to evaluations that use random assignment. Two types of activities in particular are well suited for randomization – those that involve a large number of geographical units (villages, districts, constituencies, etc.) within a particular country, and those that involve a large number of individual participants. In these cases, there are a number of ways that the "treatment" can be randomized across units, for example through a phased roll-out or by selecting which units are to be included in the project. While random assignment can be designed to include multiple countries, in practice single-country evaluations are likely to be most feasible. To illustrate, we offer several examples of how random assignment methodologies might be used in the context of party assistance programs.

<u>Example 1: Randomizing a Training Program across Individual Participants</u>

Our first example comes from a USAID-funded project in the Dominican Republic. The project provides training to political party members on a variety of topics related to conducting campaigns, building party infrastructure, and relating to constituents. The project was designed from the outset to create a control and treatment group by using a lottery to select participants. The first step was to inform party members of the up-coming training program and to request that those interested in participating apply to the

implementing partner. In this case, there were more people interested in participating than could be accommodated in the program. The implementing partners therefore used a random selection procedure to determine which applicants would be included in the training and which would be excluded (those who were excluded were informed that they would be eligible to apply to participate in the training in subsequent years, although they were not guaranteed a spot in subsequent training rounds). To insure that all parties were represented in the trainings, the implementers used a quota system for each party at the selection stage. This was important for the purposes of insuring that the intervention did not disproportionately benefit one party at the expense of others, consistent with USAID's mandate to provide party aid in a non-partisan fashion. To measure the effectiveness of the project at the individual level, a number of indicators were developed to track whether the participants learned the skills taught and implemented them at later stages. To complete the analysis, the participants' skills would be compared with the applicants who did not receive the training.

This is an example of what is called an *"oversubscription" design* for conducting a randomizing control trial. The main advantages of this design are, first, that the evaluators (or program implementers) can control who receives the treatment and can ensure that the treatment is administered through random assignment from the pool of potential participants. Second, because the kinds of people who would apply or be interested in participating in these programs will quite possibly differ from average party members or average politicians on a host of hard-to-measure psychological or motivational factors, this means that the treatment group and control group will likely be "matched" at the outset on these kinds of potentially confounding variables. The oversubscription method is therefore quite strong in terms of controlling for the kinds of *unmeasured confounders* that may distort causal inference in a non-experimental setting

Example 2: Randomizing an Institutionalization Project across Geographical Units

Our second example is a hypothetical case. In many emerging democracies, USAID funds projects designed to help parties build local offices in order to extend their reach outside of capital cities. These efforts are based on the assumption that more fully institutionalized parties will be better able to learn about and represent voters' needs and wants. To test these assumptions, a project could be designed to provide assistance to parties at the sub-national level, randomizing the intervention across geographical units. For example, in a country with 30 regions, a project would work with 10 of the 30 regions over a multi-year period to provide technical assistance on constructing district party offices and developing outreach initiatives in those regions. The 10 target regions would be chosen through a random selection process. To test for the effects of the project, a variety of data would be collected at the end of the project. For example, a survey could be conducted in both target and control districts to test whether voters: 1) have been contacted by parties; 2) believe that the parties understand and reflect their interests; and 3) feel that they understand party platforms. Differences between voters in the target and treatment area could be attributed to the project with a very high degree of certainty.

In some cases, a program cannot feasibly be withheld from different areas of a country or region, and so it will not necessarily be possible to draw treatment and control groups

simply at random. Nevertheless, randomization may still be possible with what is known as a *"phased roll-out" design*, whereby the *order* with which units or regions receive a program's intervention is randomly determined. In the above example, 5 of the regions could be chosen at random to receive the program at first, with the remaining 25 regions serving as an initial control group. Then, after some time interval, 5 additional regions could be randomly assigned to receive treatment, with the remaining 20 districts serving as controls, and so on. In this way all units eventually receive treatment, but the implementation of the program is done in such a way as to ensure that some units are able to serve as "counterfactual controls" to the treated units at each phase. This design points again to the need to think creatively at the design stage of a program's development about how appropriate control groups can be incorporated into an evaluation.

Example 3: Randomization through an "Encouragement" Design

In some instances, it may not be possible in advance to randomly select units for treatment because USAID or the implementing partner is not able to ensure that selected units will participate in the program, nor able to prevent non-selected units from attending. For example, a program may focus on ordinary citizens and consist of a series of open community workshops that attempt to instill greater understanding of, or confidence in, the party system. It is unlikely that individuals could be randomly selected to be treated in the program, since participation is neither mandatory nor necessarily limited to those individuals selected by the implementing partner. An alternative randomization strategy in this instance would be to generate a random sample of individuals in a given community who would be *encouraged* to attend the workshop with some kind of financial or other incentive; the party-related attitudes or knowledge of this encouraged group could then be compared after the treatment to a random sample of individuals who were not encouraged to attend to generate an initial "intent to treat" estimate of the effect of the program. A more nuanced estimate of program effect in the encouragement design can be obtained if the "compliance rate", i.e., the attendance among the encouragement group, can also be ascertained. In general, this kind of design is particularly useful for evaluating programs where participants and non-participants cannot *a priori* be determined by USAID or by the implementing partner.

2.2 Challenges and Limitations of Randomized Evaluations

While randomization of treatment is lauded as generating the most credible estimates of a program's effects, it should also be noted that evaluations that make use of randomization methods face a number of practical and methodological difficulties that must be overcome in order to arrive at valid causal inferences. One of the more serious, as mentioned above, is the potential for "spill-over effects", or interference between the control group and the treatment group. In the above example of a party training program, the validity of the evaluation would be undermined if, for example, the party members who participated in the training program shared their new knowledge with other party members in the control group. Such spillover of treatment, of course, is often desired by USAID and implementing partners, with programs being designed explicitly to encourage trainees to share knowledge with others in their professional or social networks who are not directly trained. Overcoming this potential bias is not impossible, however, and in fact designs can

be implemented that consist of multiple control groups, some with greater likelihood of contact with the treatment group than others, in order to make more nuanced comparisons between individuals (or other units) which were treated directly, individuals (or other units) who were treated indirectly, and individuals (or other units) who had no contact with the program whatsoever. In this way the potential exists for estimating the total effect of a given PPA activity, both on direct participants (or other units) and on units who came into contact with the program indirectly as well.

Other difficulties may arise in randomized evaluations that can be challenging to control. The behavior of treated or control individuals may change precisely because treatment was applied or withheld; these are the well-known "Hawthorne effect", where individuals alter their behavior as a result of their possible knowledge that they are being treated, and the so-called "John Henry" effect, where individuals in the control group may change their behavior or try harder to achieve some outcome because they were passed over for treatment. This suggests more generally that the somewhat artificial nature of the treatment in a typical experimental or randomized evaluation may pose challenges that need to be considered in drawing definitive conclusions about program impact. Along these lines, programs that are amenable to randomization in a particular geographic area or in a relatively narrow population may show effects that cannot be generalized to other social contexts, nor will small-scale program effects necessarily replicate what the effects of a program or intervention would look like if it were to be implemented on a larger, country-wide scale.

In some instances the limitations of randomized control trials will be so severe that other evaluation methodologies will need to be utilized. As noted above, evaluation of PPA activities needs to focus on *impacts and higher-level impacts* of projects, and not only what we have referred to as *outcomes* on individual participants, particular party members or candidates who took part in workshops or other kinds of trainings. These higher-level impacts involve potential changes to a country's party system, electoral laws, campaign activities and the post-election behavior of party elites, and the within-country "randomization of participants" framework will be of less utility in evaluating effects at this more aggregated national level. We discuss in the subsequent section how alternative impact evaluation methods, such as multi-region and multi-country longitudinal analyses, may be utilized for these purposes.

Along these lines, it is also the case that for many PPA activities, randomization of treatment assignment will be difficult and perhaps impossible to implement. This problem is most evident when random assignment would require political parties as a control group. From the practical side this is unworkable because there are too few in any country to create control and treatment groups. It is also ill-advised because USAID is bound by its bound by its own policy to support all "significant democratic parties" in countries where it funds political party assistance activities.[7] The common USAID program that seeks to encourage dialogues between party leaders can serve as an example. The goal of this type of activity is to bring together all major actors to discuss critical challenges and work

[7] The Political Party Assistance Policy states that "programs do not seek to determine election outcomes" and that "USAID programs must make a good faith effort to assist all democratic parties with equitable levels of assistance."

toward bi-partisan solutions. For example, in Indonesia USAID provided funding to IRI to support a national-level task force that discussed and advocated for electoral party reform. In such a case it would make little sense to intentionally exclude certain parties in order to construct a control group. A randomized control trial could conceivably have been implemented, perhaps via the "oversubscription" or "encouragement" designs discussed above, but given the small number of actors, the very high likelihood of spillover, and the stated interest in the higher-level policy-oriented impact from the program, such an evaluation would probably have been of limited benefit..

Another core component of many PPA programs is support for strengthening the internal capacity of parties. As part of this program, USAID (through IRI and NDI) provides assistance to parties on technical aspects of candidate recruitment and campaign development. These programs are often geared toward national-level party officials, and it would be impractical, count to policy, or unethical to exclude some parties from such programs for the sake of creating a control or comparison group. Randomization of treatment via the "phased roll-out" method is a possible means of implementation for these kinds of programs, but likely to be of relatively marginal benefit, given the costs involved and goals of the program. In short, some PPA programs simply will not be amenable to randomized evaluation procedures.

Finally, even in cases where projects *could* in principle have been evaluated via random treatment assignment as a program is implemented, it is nevertheless the case that most evaluations of PPA activities (and other USAID programs) take place after programs have already been completed, and after the opportunity for randomization of treatment has passed. We have argued throughout the document that projects should take evaluation needs into consideration much more strongly at the design and implementation phases. But even if this recommendation were to be implemented, there will still be a need for evaluating the effects of programs that were not conducted in this fashion. In the next section, we describe alternative evaluation methodologies that, to the extent that certain assumptions underlying the methods are satisfied, are still able to provide valid estimates of PPA program effects. Chapter Six of the NAS 2008 report provides additional guidance on the use of these kinds of evaluations

2.3. Evaluations with Non-Random Assignment of Treatment

All of the above considerations mean that, in many instances, it will not be feasible to evaluate a program by making use of randomized treatment assignment. Programs may in in principle be less amenable to randomizing treatment units, or programs may be implemented without random assignment of treatment built into the design. For example, NDI and IRI may choose to work in just some among a country's many regions, or USAID might support local party offices in some but not all areas of a country, or they may run training programs for women in a subset of provinces. The goal of these methods is to compare the observed outcomes (impacts) associated with the non-randomly assigned treatment units to the observed outcomes (impacts) associated with a meaningful comparison or control group, and to use a variety of specialized procedures in order to control for possible confounding factors that could bias the comparisons. The most powerful of these methodologies employs over-time or *longitudinal* comparisons of the

(non-randomly assigned) treatment and control groups, a design which may be used in a variety of single-country and multi-country settings to assess program outcomes as well as impacts and higher-level impacts. Single-shot observational comparisons of treated and untreated units are also possible, though with greater possibilities of "hidden biases" that may confound the causal inference process. We discuss in this section various quasi-experimental designs for program evaluation, and conclude with suggestions for evaluation when the conditions necessary for rigorous designs of this type are not present.

2.3.1 Longitudinal Designs

Consider a PPA program designed to promote the success of women candidates for local and national office in which the implementing partner sponsored training sessions in some regions of a country but not others. The regions were chosen for a variety of technical reasons, for example the availability of appropriate infrastructure, ability and willingness of the partner to operate in certain regions, and pre-existing levels of female literacy rates and other gender-related resources. How could the effects of this kind of program be ascertained?

Clearly a simple comparison between the number of women candidates or their electoral success in "treated" regions versus "untreated" regions would be a biased estimate of program impact. This is due mainly to the problem of "selection bias," as the treatment and control units differ on a host of factors that, aside from the treatment, could impinge either positively or negatively on women's electoral success. Some of these factors could be measured and included as statistical controls in an effort to isolate the effects of the program, but it is likely that many pre-existing differences between the treatment and control areas would be very difficult to measure and include in the analysis. In formal terms, the treatment and control group would still not be identical in all ways aside from treatment exposure, once additional observable factors were taken into account, and hence the control group's level on some outcome or impact would still not be a meaningful "counterfactual" proxy for what the treatment group would have looked like in the absence of treatment. However, the evaluation can be strengthened considerably by collecting data on outcomes and impacts for both the treatment and control groups over time, that is, before and after the program is implemented in treatment areas. With two "waves" of observation, pre-and post-treatment, the evaluation then becomes a comparison of the difference that the treatment group exhibits on some outcome or impact with the difference that the control group exhibits over the same time interval.

This "difference-in-difference" design has several important advantages. First, to the extent that unobservable factors on which the treatment and control groups may differ are *stable* over time – as they may be here in the case of relevant factors such as "regional political culture" or "female empowerment" – their potentially confounding effects will be eliminated through this design's differencing procedures. And because of this, the assumption necessary for this design to provide unbiased causal effects of a program is, relatively speaking, easier to justify than in designs without longitudinal observations. In this design, we need only assume that the *rate of change* that the control group exhibited on some outcome or income would have been the same *rate of change* that the treatment group would have exhibited in the absence of treatment. This points to the critical

importance of gathering baseline data on desired outcomes and impacts as part of the evaluation process; with pre-and post program data, the difference-in-difference design can be a powerful method for controlling stable unobservable confounding factors between the treatment and control group, and thus arriving at unbiased estimates of how a given program may have contributed to differential change among treated compared to control units.

Longitudinal designs that are conducted may be especially useful in assessing *impacts* and *higher-level* impacts from PPA activities. Within a particular country, the method allows the comparison of party-system or party-level impacts by region, so that an evaluation could determine whether the changes in important impacts were greater in areas where PPA activities were present compared to areas where they were absent. Multi-country evaluations can also follow this general strategy in assessing whether PPA interventions that were non-randomly allocated affect *country-level* impacts and higher-level impacts related to the parties or party systems. For example, this approach could be employed to examine whether USAID funding at the sub-sector level correlates with changes in the degree of party system institutionalization, volatility, nationalization, female representation, or other sub-sector-level indicators over time.

The longitudinal approach can also be extended in several useful ways. First, the "inputs" in the analysis need not be a simple dichotomous indicator of whether a region or country was a "treatment" or "control" unit. It is also possible to analyze the *amount* of PPA funding that a given unit received, so that, for example, countries can be compared longitudinally in terms of the gains/losses they exhibit on important impacts and higher-level impacts as a result of the total level of party-oriented assistance they were provided. This approach was used, for example, in a recent evaluation that examined whether total USAID funding leads to increases in a country's overall level of democracy, and that study also included similar kinds of subsectoral analyses that could be directly applied in the party strengthening.[8] Moreover, longitudinal designs could (and should) include more "waves of observation" that a simple two-wave "pre-treatment/post-treatment" set-up. With multiple waves of observation, the evaluation can make considerable progress in ruling out one of the main challenges this design faces in making successful causal inferences. If, for example, the areas where USAID or implementing partners choose to "treat" were those areas that were already showing a more positive or negative underlying trend on some party-level impact, the two-wave pre-post design would fail to distinguish the effect of the treatment from the effect of the underlying trend. (Technically, the control group's pre-post difference would no longer be a valid counterfactual for the treatment group's pre-post difference in the absence of treatment). With multiple waves of observation, the analysis can include both a treatment indicator (or level of effort variable) along with a measure of the treatment trend (or time-related effect) on the impact/higher-level impact that may be different for treated and untreated units. This kind of analysis, then, could provide information on how units/areas/countries may have changed as a result of PPA funding, over and above how

[8] Finkel, Steven E., Aníbal Pérez-Liñán, and Mitchell A. Seligson. 2007. "The Effects of U.S. Foreign Assistance on Democracy Building, 1990-2003." *World Politics* 59 (3): 404-39

the non-randomly assigned treatment units may already have been changing differentially from the control units.

The challenge with this strategy is that it typically requires considerable data collection efforts. This impact evaluation approach, like random assignment, makes serious data demands for evaluation, and should be integrated with program design from the beginning. For single-country evaluations, USAID should be attentive to opportunities to leverage within-country variation for evaluations, particularly when projects are implemented in some regions or districts but not others. Moreover, the same kinds of potential problems discussed above related to "spillover" and "interference" effects, and the possible alteration of treatment and control group's behavior based on their treatment status, may also occur in evaluations with non-random assignment of treatment. Most importantly, non-random assignment means that we cannot rule out completely the possibility that unobserved factors that are not stable were also related to both treatment assignment and the outcomes/impacts/higher-level impacts under investigation. If, for example, USAID conducts consensus-building workshops in particular areas at a given time because of some idiosyncratic factors that made those areas especially "ripe" for consensus at that time, the estimate of program impact would be biased. More complex statistical procedures exist to make progress in ruling out these kinds of confounding processes, but they require additional waves of observation and their own sets of assumptions that must be justified in a given case.

2.3.2 Cross-Sectional Evaluation Designs

More challenging are evaluations in which non-randomly assigned treatment and control units are observed at a single point in time. For example, baseline (pre-treatment) data may not be available for individuals who were trained in a PPA campaign practices program, nor available for the individuals in the control group who were not trained, so the evaluation may involve a single shot post-treatment comparison of the two groups. In these cases, the potential for selection bias is especially severe, as "stable unobservables" that could confound the analysis cannot be differenced out, nor can multiple waves of observation take differential underlying trends between the treatment and control groups into account. Nevertheless, rigorous methodologies for these kinds of evaluations are available, taking one of two general forms.

The first strategy involves "matching" treatment and control units on a series of observed factors (or "covariates") that distinguish the treatment and control groups, calculating the difference in some outcome (impact) within the matched strata, and aggregating these differences into an overall "treatment effect." Various matching methods are utilized, with the most prominent being a general procedure known as "propensity score matching," whereby the analyst develops a model that predicts the likelihood ("propensity") that units will receive treatment from all known covariates, and then matches each treated unit with the control unit that had the most similar propensity to be treated. In this way the evaluation estimates the difference on outcomes between treated units and the "most similar" control units which serve as the proxy for what the treated units would have looked like in the absence of treatment

These kinds of evaluations are widely utilized in other development areas and could readily be applied to PPA project evaluations as well. The main challenge to successful causal inference here is in this design's inability to match on *unobservables* – things that may differentiate treated and control units that are difficult to bring into the analysis but which may be highly relevant to a unit's propensity to experience a given treatment. In cases where program participation is the result of *self-selection* into a treatment group, unobservables related to motivation or personality characteristics may be especially relevant. The evaluation will need to attempt to develop ways of taking these factors into account, either through observed variables that can serve as partial proxies for the unobservable characteristics, or through sensitivity analyses that can provide information on how treatment effects may change, depending on different assumptions about the magnitude and direction of the unobservables' effect on treatment status and on the outcomes (impacts) under investigation.

The second evaluation strategy involves including additional information in the form of what are known as "instrumental variables" into the analysis in order to identify a program's effects. Consider a program designed to increase voters' awareness of the issues and party programs during a particular election campaign, where individuals self-select into the treatment and hence where unobservable differences between treated and untreated individuals will be difficult to control. The basic idea here is to find a variable that is a) related to treatment status, b) likely to be unrelated to the confounding unobservables, and c) unrelated to the outcome (impact) under investigation except through its effect on the treatment itself. If such a variable can be found, it can serve as an "exogenous proxy" or "instrument" for the treatment; effectively, it substitutes for treatment status and, to the extent that the assumptions of the method hold, its estimated effect provides an unbiased estimate of the program's treatment effect.

In the current example, it may be that training workshops are held in a variety of locations in a given area, and that an individual's likelihood of attending the workshop is a function of how far he or she lives from the training locale. To the extent that the distance an individual lives from the training locale is unrelated to his or her knowledge about political parties or campaign issues, and to the extent that the distance an individual lives from the training locale is unrelated to personality, motivation or other unobservables that could relate to party knowledge, then *distance from the training locale* may be a valid "instrument" for treatment status, and the effect seen from this variable in an instrumental variable (or "two-stage least squares") analysis would yield un unbiased estimate of the program's impact. Evaluations making use of instrumental variables methods are becoming increasingly common in assessing interventions in developmental economics, public health and educational programs, and we see utility in their application in the democracy and PPA sectors as well. Still, the assumptions underlying use of instrumental variables methods are relatively strict, and appropriate instruments will often be very difficult to find and include into the analysis. As always, creative thinking about possible instruments should be undertaken at a program's design phase.

2.4. Performance Evaluations

It will not always be possible to evaluate programs using the formal impact evaluation methods we have presented thus far. In many instances, evaluations will be commissioned well "after the fact" – i.e., at some time after the program has already been implemented. When this happens, it may not be possible to construct valid control groups and in some cases it will be impossible to assemble time series data that stretches back to the beginning of the project. Given these constraints, the difficulty of disentangling the effects of USAID's programs from other factors will be most acute in this setting. Here we describe some of the problems associated with these evaluation approaches, but also explain how to improve them. We suggest, first, that evaluation teams attempt, to the extent possible, to construct comparative groups by collecting information about the pre-treatment period or about individuals or regions that did not participate in the programs. We then describe an analytical approach to these evaluations—which are most effective if there is a control group but still valuable in their absence—that has two distinct but complementary components. The first component starts with program activities and seeks specific evidence to determine direct results from the activities. Because starting with activities may not provide much information about progress towards development—which provides the rationale for the projects—the second component asks evaluators to enumerate the role of parties and party systems within the context of a developed democracy, look for evidence of changes, consider the range of factors that might influence these goals, and then seek links between the changes and program activities. In Appendices 1-4 we provide specific guidance for carrying out these components of the evaluation.

In attempting to relate movement in impact-level indicators to USAID programs without the aid of pre-designed comparison groups, evaluators conducting performance evaluations will have a special burden of considering the multitude of factors that affect party and party system change in particular contexts. Parties and party systems are influenced by political challenges and exigencies, the institutional framework and changes thereto, and new ideas that they may learn from development programs or elsewhere. Evaluators, therefore, must identify these factors for each goal under study, with the challenge of measuring the relative impact—or at least the directional impact—of each. As noted, USAID programs are just one factor impacting on parties, and it is not possible to know with certainly how parties would react in the absence of these programs. These evaluations, therefore, may only be able to determine if the programs are working in the direction of trends, or trying to counterbalance negative trends. It is important to note that in some cases a lack of progress on indicators should not be interpreted to mean that a given USAID program is not having the desired effect. It is possible that without the program, conditions would be worse and that maintaining the status quo should be interpreted as evidence of success. This is likely to be the case when confounding factors are pushing in the opposite direction of USAID's programs. Thus, in some cases USAID will want to continue funding programs even if the relevant impact indicators show no positive movement over time.

In addition to the in-country design we have discussed, performance evaluations may also be conducted as cross-country comparisons. Small-N multi-country evaluations typically

include a few countries that are purposefully chosen to facilitate the study of a particular question. This approach might be used, for example, to help understand why a specific project type worked well in some countries but not in others, in order to improve USAID's knowledge of how contextual factors affect the success of a given project type. If USAID were interested in understanding, for example, why projects designed to help parties become more fully institutionalized succeeded in some countries but failed in others, a natural approach would be to conduct an evaluation that pairs success cases with failure cases.

It is important to note that the traditional performance analysis that relies heavily on interviews with party members, NGO leaders, academic analysts, and politicians are not without value. Evaluators, for example, can learn whether program participants enjoyed the program and can perhaps ascertain what the participants learned and how they put their new knowledge into practices. They may also learn of criticisms, especially when the interviewee compares training programs run by USAID-funded groups with different organizations.

While these and other lessons are informative, they cannot provide conclusions that can withstand rigorous scrutiny. Analytically rigorous evaluations are complex and perhaps costly, but they are the only way to generate conclusive evidence that ties programming to outcomes. But despite the problems, performance evaluations are common. In this section, therefore, we provide suggestions to build on the strengths of the traditional performance design in ways that add rigor to the study. We provide a system of structured inquiry that promotes comparisons where possible, but also suggests data that collectively can help build circumstantial evidence about the value of programs.

2.4.1 Data Collection and Analytic Procedures for Performance Evaluations

As noted above, we suggest that the evaluation include two distinct but complementary components. The first component focuses on activities and seeks to determine whether there is any evidence that USAID-funded activities produced positive results. At this stage, the evaluators will primarily be concerned with effects on outcome indicators that are directly linked and proximate to the activities. The second component starts from the party development characteristics described in the Conceptual Framework (accountability, good governance, etc), and seeks to determine whether the country has made progress toward these aspects of development. In practice, the data collection process for these two components will overlap to a considerable degree. It is valuable, however, to distinguish between these two aspects of the evaluation for analytic purposes. In short, the first component looks for direct effects of activities while the second looks for larger changes at the sub-sector level.

In conducting the study, evaluators should work to find data on comparisons groups, as in the experimental and quasi-experimental designs discussed above. For some projects, this will imply collecting data about the pre-program period, in order to assess progress on program or development goals in relation to the timing of programs. For other pieces of the analysis, evaluators can add validity to their findings by collecting and comparing data in regions that have and have not benefited from the program . If programs treated

individuals, then analysis teams should attempt to define a comparative group of similar but untreated persons.

- *Component 1: Evaluating the Effectiveness of Activities*

For the first component, evaluators will seek to match program activities with expected outcomes. To facilitate and systematize this process, Table 3 suggests detailing all activities for the analysis and then identifying the outputs, outcomes, and outcomes for each activity. In this exercise, the evaluators would also have to identify indicators for each of these—preferably ones that are available for over-time or cross-sectional analysis. Appendix 1 provides a table that lists several dozen objectives for the typical conferences and trainings (activities) and provides examples of the associated outputs, outcomes, and "higher level" impacts as defined in the *Conceptual Framework* and CEPPS. The final column lists examples of qualitative or quantitative data needed for analysis. Note that the appendix tables list only a limited number of activities for the "Organizational and Technical Capacity" category, as most of the related themes are incorporated into the other development categories. This is intended to encourage the emphasis on how the programs ultimately affect democratic development.

Table 3: Working Table for By-Activity Analysis

Activity	Output	Outcome	Impact	Higher Level Impacts
Accountability Representation and Participation				
1 2 etc.				
Governability and Good Governance				
1 2 etc.				
Stable and Peaceful Contestation"				
1 2 etc.				

*Because it is just an illustration, the table does not include other rows for "Rule of Law and Fair and Honest Elections" and "Organizational and Technical Capacity." As suggested in the appendix, the table could also be divided for programs aimed at partisan actors, programs aimed at institutions and institutional reform, and programs for non-partisan actors. .

The tables, in sum, are meant to encourage a rigorous analysis by enumerating the programs and carefully explaining the indicators for each. As we have explained, this type of analysis will be most compelling if the team is able to provide information about baselines about comparison groups. Especially when this is not possible, the team should be aware that the significance of confounds rises with the level of analysis (from output to outcome to impact). Still, it may be useful to speculate about the larger effects of the

programs, since they are the stated purpose of the programs. In so doing, however, it will be necessary to discuss the many contextual (i.e. confounding) factors beyond the programs that affect party and party system development.

- *Component 2: Tracking Progress towards Party Development*

The second component should focus on the levels of party and party system development, , and work to associate programs with indicators of changes over time or differences in treatment and non-treatment groups . This component of the evaluation will detail development characteristics, ask if they were foci of the programs, and then collect statistical or interview and other types of data to assess effects. Collecting information without a defined comparison may be informative, but especially if there are no clear control groups, the analysts would only be able to determine tentative relationships about the effects (or lack thereof) of program activities. Still, this type of analysis is crucial to promoting an analytical focus on the larger goals of the programs. Table 4 provides a format for collecting the relevant data and Appendices 2A and 2B supplement this table with a detailed list of the development characteristics and the data necessary to evaluate them.[9]

Table 4: Working Table for By-Development Characteristic Analysis*

Development Characteristic	Goal Identified in RFA or Work Plan? (y/n + details)	Level of Priority/ Effort Explanation	Activities Designed to Address	Data Needs for Evaluation	Control Group (Time series or Cross-Sectional)
	Accountability, Representation, and Participation				
1 **2** **.** **.**					
	Governability and Good Governance				
1 **2** **.** **.**					

*Because it is just an illustration, the table does not include rows for "Stable and Peaceful Contestation" and "Rule of Law and Fair and Honest Elections." The Appendix also suggests separating development charaterisitics of parties from those pertinent to party systems.

[9] Unlike the "by-activities" tables, these tables exclude the "Organizational and Technical Capacity" category, because it is defined here as a means, to democratic development not an ends.

Table 4 and Appendices 2A and 2B are meant to suggest that evaluation teams first consider which aspects of development are foci for the target country, thinking of these in terms of the three parts of development goals for parties and party systems. Many of these goals are discussed within the Conceptual Framework, and the table is divided into the corresponding broad characteristics of development . The first column of table asks the analyst to list the specific PPA goals for programs around the world (see Appendices 2A and 2B for an extensive list), and in the second column the team is supposed to identify which of these programs were applicable to the country under study.[10] The third column then asks the team to assess the level of effort expended in addressing the goal. Was this a central or minor part of the project? What were the costs involved? The final column asks the evaluators to identify activities directed towards these goals, which will help to tie the two components of the evaluation together. Once that step is completed, the team will shift to their main effort: collecting data on the change in the indicators over time or in comparison with other groups or regions. The appendix tables also provide suggestions for indicators of each of the more than 50 goals we have identified,. This list, still, is not comprehensive and there are surely more (and better) indicators for some goals. The purpose of providing these tables, however, is to encourage evaluation teams to list all development characteristics, define and measure specific indicators for them, and consider longitudinal or cross-sectional comparative groups if possible. To reiterate, if there are no control groups this part of the evaluation cannot provide conclusive evidence about the (positive or negative) effect of programs, because of the many confounding factors that influence these goals. The evaluation, therefore, should include a detailed discussion of the confounding factors and whether they work with or against the programs. By combining this review with evidence about the changes in development goals and emphases of programs, this exercise can serve as a central component of a comprehensive evaluation.

2.4.2 Suggestions for Improvement in Performance Evaluations

As noted earlier, the evaluation model where external consultants are contracted to evaluate a recently completed program by spending a relatively short amount of time conducting in-country interviews and gathering data is relatively poorly suited for impact evaluations. Such evaluations may be able to provide suggestive evidence regarding the effects of USAID programs, but they will typically be unable to document USAID's impacts with a high degree of certainty. To improve the value of these evaluations, we offer five suggestions.

First, USAID should place greater emphasis on collecting data that can be used to measure key trends in outcome and impact variables. For example, USAID frequently supports projects designed to increase the number of women who serve in party positions and who hold elected office. Yet, program implementers have not been required to systematically collect relevant data on these outcomes over time, making it difficult for evaluators to track progress toward the project's core goals. For future programs of this type, USAID ought to

[10] These goals are set out in the CEPPS agreement (Appendix Table 3), as well as the more specific goals set out for the particular country which are documented in publicly available RFAs (Requests for Applications), PDs (Project Descriptions, non-public applications for CEPPS programs, Performance Management Plans (PMPs), and work plans that are required from the award recipients.

place greater emphasis on constructing baselines and collecting trend data, upon which external evaluators could draw in assessing the effectiveness of such programs. Even in cases where it will be difficult to attribute credit for positive (or negative) trends to USAID, establishing whether progress has been made toward broad party and party system goals is an essential starting point for efforts to examine program effectiveness.

The second and related suggestion is that evaluation teams should seek to define and collect data from non-treated individuals or regions to serve as a comparative group. The traditional studies emphasize discussions with stakeholders and individuals who have received training from USAID partners. To validate a program for female candidates, for example, the teams should collect data about women in regions where there were no training programs as well as where there were programs. This would help to determine the program's influence in preparing women for elections or leading to their successful campaigns.

Third, in practice evaluators sent to conduct a short-term performance evaluation will be required to make a series of choices about what to focus on during the country visit. Specifically, evaluators will need to decide whether to evaluate the overall CEPPS program within a country or to focus on particular implementers, projects, or activities. Developing indicators for the overall success of the program, however, will be necessarily more abstract than indicators of specific activities. Further, it will be even less likely that evaluations of the full program will have a valid control group, and as such these types of analyses cannot produce conclusive results. Further, given the time constraints, it will be unrealistic for the evaluation team to try to assess multiple projects implemented by different actors in different parts of the country. Prior to departure, therefore, evaluators should decide (in consultation with USAID and the program implementers) on a limited set of components to examine during a country visit.

Fourth, in practice it can be difficult to delineate between retrospective evaluations and forward-looking assessments, because evaluators necessarily consider the types of PPA projects and programs that would be appropriate for particular settings. The overlapping work suggests that the teams should work contemporaneously and with one another.

Fifth, USAID ought to encourage its partners to standardize the language and reporting systems in their reports. In the final reports for Indonesia, for example, after brief summary and background sections, the NDI report is organized around a hierarchy that begins with program goals and then moves to objectives, while the IRI report starts with program activities and then moves to "components." These inconsistencies complicate the evaluators' task, because the evaluation has to focus at a particular level of analysis. If the terms "goals," "components," "activities," or "programs" have different meanings for the involved agencies, the evaluation team will have difficulty in creating a focused and comparative study.

Part 3. Summary and Recommendations

While political party aid creates some special evaluation challenges, careful attention to methodology can greatly improve the quality of information that an evaluation provides. Our emphasis in this document has been to encourage teams to consider control groups from which to compare those who received the benefit of PPA activities from those who did not. Random assignment provides the most valid type of control group, but longitudinal and cross-sectional analyses are reasonable approximations. We emphasized that even performance evaluations can collect and productively use qualitative or quantitative data about control groups. To do so, however, requires careful pre-trip planning and decisions about which specific aspects of a program the team will evaluate. Careful planning is not always enough, however. If the political context changes, so too will programming priorities. In these cases programs may need new or baseline data than what was originally collected. Of course, the desire for rigorous evaluation cannot completely drive the programming.

One other emphasis with particular validity for performance evaluations was a consideration of confounding factors. Randomized evaluations avoid this problem, but we recognize that PPA programs do not always lend themselves to this technique. In other cases, evaluation teams must pay special attention to how factors other than the programs contribute to or hinder party development.

Control groups and a consideration of confounding factors, in sum, are necessary to help evaluators determine with a high degree of confidence whether USAID's funding has contributed to desired outcomes. When successes are achieved, USAID may be able to claim some credit for contributing to a positive outcome, but without a carefully designed study, it will not be possible to separate the effect of USAID's effort from other international and domestic actors pushing for the same results.

We also want to re-emphasize that evaluations can serve at least two purposes: informing about the utility of a particular program or about the progress of a party or party system towards a higher-level development goal. Implementers have more responsibility for the former, but evaluations at this level may be less informative about the latter.

We conclude with several recommendations designed to improve evaluation practices of political party assistance.

- USAID should have well-defined outcome, impact, and higher-level impact variables that are expected to be affected by a given PPA programs in advance of actual implementation. Without greater attention to defining the specific outcomes (impacts) that a program is intended to influence, rigorous evaluation is next to impossible. The Conceptual Framework from this project is designed to assist in this effort, along with the Appendices to this document.
- Along these lines, USAID must place greater emphasis on collecting high quality data that can be used to measure key trends in the intended outcome and impact variables. For example, greater use of national and regional surveys of the

population of interest before and after program implementation (regardless of whether the program's treatments were randomly assigned) can provide evaluators with the kind of baseline and follow-up data from which to assess the impact of the intervention.

- Evaluations must include observations on "control groups" that were unaffected by PPA interventions, and not simply rely on observation of units (individuals or regions or parties) that were "treated." High-quality evaluations, depend on credible "counterfactuals,", i.e. proxies for what the treated units would have looked like in the absence of treatment. This consideration is perhaps the most fundamental tenet of the evaluation process, and bears repeated emphasis, given that most energy, resources and attention at USAID are (understandably) devoted to those units that *are* selected for treatment. Without sufficient attention to *untreated* units, though, it will not be possible to undertake rigorous assessments of program impact.
- More generally, USAID should design Political Party Assistance programs while taking into account evaluation considerations to a much greater extent than is currently the norm. This includes, but is not limited to, constructing control groups using random assignment of treatment by taking advantage of different randomization strategies (i.e. oversubscription, phase-in, encouragement design), collecting pre- and post-intervention data for treated and untreated units when randomization is not feasible or unethical, and thinking carefully about possible "instrumental variables" that may be used in subsequent analyses of program impact.
- Evaluations must be sensitive to the needs for balancing methodological validity with judging higher-level impacts. Randomized control trials, for example, may provide credible estimations of a program's immediate *outcomes*, but may be less useful in assessing higher-level impacts related to a country's party system characteristics or most generally on levels of democratic development. This suggests that a mixed strategy for impact evaluation will be most useful, as different methodologies will contribute in different ways to assessing the full panoply of possible effects of PPA activities.

Appendix 1: Worksheet for Performance Evaluations, by Activity

----------Indicators----------					
Objective of Activity	**Output**	**Outcome**	**Impact**	**Higher Level Impacts**	**Special Data Needs***
Accountability, Representation, and Participation					
Programs aimed at Partisan Actors					
Campaigns that focus on policy issues and strategies	Number of workshops on value, use of data, and message development process	Specific policy-based messages developed and approved by party leaders	Public recognition of message topics	• Increased voter knowledge of party positions • Increased retrospective voting	population survey of party ideology/policies
Parties develop legislative agendas	# workshops on value, agenda development	Legislative agenda developed and approved by party leadership	Agenda guides party policy positions, messaging, and caucus activity	Voters identify policy positions of party	• Party platforms and other policy documents; • surveys of voter information about party positions
Concern for responding to local-level citizen concerns (constituency service)	Strategic plan identifying constituent geographies, population groups, resource commitment	Implementation of plan, e.g. district offices, constituent service systems, website, newsletter, SMS, staff	Increased contact between party officials and constituents, e.g. letters, electronic messaging, walk-ins;	Voters recognize parties as being attentive to their concerns	• Population survey about contact with representatives, by region • Records about constituency service
Improved communication with voters /3	Number of parties trained; number of participants/ party	Parties create systems (committees, regional offices, public meetings) to receive information from and respond to voters	Parties act on information received from voters to change platforms or policies	• Citizens increase communication with parties; • More policy and service requests • More citizen trust of parties	• Number of regional party offices; • Party organigram • party platforms and changes; • surveys about trust and support of parties
Development of	# workshops on	Parties improved	Strategy developed and	Population learns	• Surveys of parties'

communication strategy	development process	relations	leaders		• Population survey about parties' policies
Increased participation of members in party decision making	Party leaders and others participating in discussions about internal party democracy	Parties modify by-laws to formalize wide participation in decisionmaking	• Low and mid-level party leaders increase influence in party • Turnover in party leadership	• Increased diversity in party leadership • Stable support, even after change in leadership	• Party by-laws • Interviews about de facto vs de jure decisionmaking • Turnover of party leadership
Better organization of party congress or other events	Planning meeting	Congress/event takes place	• Leaders chosen or policies defined at event	• Membership approves of process;	• Interviews/surveys of party members (regional and of different parties), dependent on holding events
Strengthen membership outreach among women/disad-vantaged groups	Strategic plan developed/ approved	• Targeted party events • Directed advertising campaigns •	Membership in identified populations	Diversity improves in party and political leadership	Sample of membership data, by region
Candidate training for women, youth, and underrepresente d groups	Strategic plan developed/ approved	• Implementation of plan, e.g. court civil society leaders for input • Perhaps promotion of quotas or changes to electoral laws	Increased diversity in candidacies	Increased diversity in regional or national legislatures (or executive positions)	Gender, age, and ethnic composition of legislatures nationally and in different regions
Recruitment of women & disadvantaged groups to party leadership positions	Strategic plan developed/ approved	• Court local notables & NGO leaders & receive their recommendations	• Increased diversity in party leadership	Increased diversity in politics generally	• Gender composition of party leadership nationally and in different regions
Training of candidates on public speaking	# people trained	• Improved confidence and knowledge of candidates	• Increased quality & # of speeches	Trainees do well in elections	• Surveys of candidate knowledge • Election results
Outreach to interest groups and civil society	Attendance at training sessions or conferences	• Strategic plan for outreach developed, approved, & Implemented	On going interactions between parties and civil society groups	• Groups policy positions addressed • Improved policy	• Interviews with civil society groups • Legislative and

				focus of parties and campaigns	campaign agendas
Use of candidate choice procedures (including primaries) for choosing candidates or party leaders	# workshops on value, methods for instituting internal democracy processes	• Use of congresses or open primaries • Allow input from subnatl party offices to rank candidates, • Internal elections for selection committee	• Diversity among, , change in party organizational leadership; • Approval of party leadership among party members	Members and voters perceive increased democracy	• Party rules • Survey of party members (national and regional branches)
Development of (and better communication with) party branch offices /6	Number of workshops on value, process for identifying needed communication paths	• Strategic plan developed • Priority locations identified, •	• Personnel & financial resource committed; • Sub-national offices set up	Increased voter contact with and knowledge of party activities and positions of parties	Survey of party members regarding contact with parties and knowledge of party positions and activities
Construction of policy-based platforms	# workshops on value ,use of data, and platform development process	Parties use survey or other data to develop platform	Public recognition of platform topics	• Higher approval of parties as institutions • More retrospective voting	Population survey of party ideology/policies
Promotion of public debates among parties	Parties plan or agree to participate in candidate debates	Debates held (candidates for executive or legislative office; national or regional)	Parties develop policy positions	• Improved focus on policies • Public learns of party positions	• Viewer data • Media coverage of debates • Surveys of voter knowledge
Polling to provide parties with information about voter concerns	Discussions of survey results	Parties incorporate survey results into platforms, campaigns, and legislative agenda	Parties develop own survey capacity	Voters more likely to respond that parties respond to citizen concerns	• Interviews and other analysis to show relation of citizen concerns to platforms or campaigns • Surveys showing whether citizens see parties as responsive to their concerns
Parties sign and adhere to codes	Number of parties participating in	Parties sign code	Campaigns emphasize policy	Voters learn policy positions of parties	• Expert analysis of policy focus in

of conduct that emphasize campaigning based on policy debates	multi-party conversations about codes of conduct				campaign • Survey of voter knowledge of party positions • Review of party campaign literature
Improved get-out-the-vote campaigns	Strategic plan developed/ approved	Media campaigns, door-to-door, SMS, precinct info booths	Increased turnout for parties (and in regions) that implemented plan	• Increased turnout generally • Better informed voters	Turnout rates, over time and by region; electoral results
Parties encourage voters to mobilize peacefully	Number of parties and NGO trained; number of participants Types of groups represented (and not represented)	Parties and NGOs communicate with supporters and plan peaceful demonstrations	Parties and NGOs hold marches and electoral rallies, which continue to support the democratic process	• Voters learn about party positions and increase perception of system efficacy • Outsiders and populists are discouraged	Surveys testing voter views about participation in politics and efficacy of the system
Programs aimed at Institutions and Institutional Reform					
Electoral law reform to • improve ties between voters, parties, and legislators • increase diversity choices among parties • ensure that the number of parties balances representation & governability • consider how the range of parties	Participation in conferences and fora on electoral systems, party finance laws, primaries, executive-legislative relations and other aspects of institutional structure	Parties pursue talks about institutional reform	Laws proposed	New laws passed	• Electoral law analysis Voters' use of preference voting (if possible) • Number (effective) of parties • Survey to show ideological range of parties • Electoral data to show party nationalization & ethnic bases of party support • Surveys to assess factors impacting party support (ethnicity, policy position, candidate qualities)

balances regional or ascriptive identification with nationally-oriented catch-all parties					
Consideration of quotas to encourage increased participation of women and other underrepresented groups	Meetings with NGOs and parties	NGOs or parties hold fora to consider quotas or other reforms Public and media support for reform	Reform proposed and passed in legislature	Diversity improves	Number of women or people from disadvantaged groups in national or provincial legislatures
Support for use of and recording of legislative roll-call votes and other legislative activities	Meetings with parties about legislative transparency	Legislature debates use of roll call voting	Legislature revises procedures for roll call voting	• Legislature publishes roll call voting results • Media and watch-dog groups begin reporting roll call results	Number of roll call votes taken Legislative rules
Programming Aimed ad Non-Partisan Actors					
Enhanced voter information about party and government activities	Number of voters participating Reach of advertising	Parties support information campaigns Parties publish information for citizens NGOs publish and distribute information for voters		Improved voter knowledge of party and government activities	Surveys of voters
Non-partisan groups promote inter-party	NGOs organize debates	Debates held, at national or regional levels	Candidates develop policy positions	Voters learn parties' policy stances	Survey to test voters' knowledge of party positions

debates					Information about viewership & media coverage of debates
Governability and Good Governance					
Programming Aimed at Partisan Actors					
Training for legislators and parties on budget and bill analysis	Number of people trained	Parties dedicate personnel and resources to budget and bill analysis	Parties propose bills and substantive amendments	• Legislative bills and amendments pass with more frequency • Increased policy focus of parties	Bills proposed by legislature and executive
Facilitating coalition and/or consensus building for policy and institutional reforms	Parties participating in private and meetings &multiparty fora	Parties initiate multiparty conversations on policies or coalitions	Multiparty legislative caucuses formed	Electoral and cabinet coalitions endure based on shared policy positions	• Roll call voting analysis to show which parties supported important legislation • Makeup and activities of caucuses
Facilitation of negotiations among parties for cabinet formation and maintenance	Parties hold joint meetings to discuss policy convergence; # meetings with parties to discuss cabinet or electoral alliances	Parties hold internal debates on merits or policy, electoral, or cabinet coalitions	Parties co-sign legislation, form multi-party cabinet coalitions, or join in electoral coalitions	Approval of policy by multiparty agreements; reduced number of parties	• Legislative voting records on budget or important legislation • cabinet composition • diagram of party coalitions
Programming Aimed at Institutions and Institutional Reform					
Conferences to consider institutional reforms that affect legislative powers and hence coalitions	Participation of parties in conferences	Commissions formed to consider institutional reform	Reforms proposed	Reforms approved	Indices of executive and legislative powers
Facilitating discussions with the executive	Participation by representatives of executive and the	Reforms proposed and implemented	De facto practices modified	• Media reports on and uses improved access to	• Legislature accessing information • Interpolation of

branch to improve parties' and legislatures' access to information, thus facilitating oversight	parties in meetings			information • Legislature develops oversight system that makes use of new access	ministers • NGO and media reports based on new information sources
Building a constructive opposition	Parties and legislative leaders participate in meetings to support opposition influence	• Naming shadow government • strengthening committee policy expertise • review of committee rules of order • improved staff resources for opposition	• Legislative opposition proposes bills and amendments • Improved legislative oversight of executive and policy process	• Increased policy focus of parties • Increased retrospective voting	• Bills proposed by parties in opposition • Population survey of party ideology/policies for analysis of retrospective voting
Parliamentary caucuses	Meetings to identify potential caucuses, their resources, organization, and legislative agenda	• New caucuses formed, • Legislative resources dedicated	Caucus functions as a coherent group in pursuit of its agenda	• Increased diversity in legislative leadership • Improved policy orientation • Increased legislative influence in policy process	• Source of legislation proposed and passed • Composition and activities of caucuses
Stable and Peaceful Patterns of Competition					
Programming Aimed at Partisan Actors					
Development of a Code of Electoral Conduct	# workshops on value, code content, communication/adherence strategy	Code and strategy developed and approved by party leadership	Parties focus campaigns on issues and promote democratic process	• Increased approval of parties as institutions • Public awareness of code	Voter surveys asking about code and party conduct
Programming Aimed at Institutions and Institutional Reform					
Facilitating discussions of party and electoral laws to consider sources of electoral	Number of conferences and participants	Parties and civil groups continue debates about alternative electoral systems and their effects; Publication of proposal or recommendations	• Public debates on party law reform • reforms enacted	Changes in party system in terms of number of parties or other concerns that motivated reform process	• Electoral volatility • Effective number of parties

volatility					
Rule of Law & Fair and Honest Elections					
Programming Aimed at Partisan Actors					
Parties mobilize (and train) volunteers for poll-watching	Meetings to plan strategy for identifying volunteers & activities	• Training events for volunteers, • Coordination with sub-national party officials	• Decreased fraud • validation of electoral results	Increased voter faith in electoral process	• Number of polls covered • Interviews with parties about faith in electoral process • reports from international election monitors • voter surveys about faith in elections
Assisting parties in strategies for raising funds	Number of parties and people trained	Parties revise fund raising strategies	Parties pursue transparent finance practices Funding sources reported to authorities	Watchdog groups report (favorably) on funding process	Reports from election authority
Training about the use of exit polls or quick-count techniques	Number of parties trained	Parties recruit sufficient numbers for analyses Parties train their own members and volunteers	• Quick-counts validate election results • Parties analyze exit polls	• Decreased accusations of electoral fraud • Parties increase understanding of electoral support	Number of polls covered by parties
Support for parties in organizing internal elections	Number of parties supported	Parties hold conventions or primaries	Turnover in party leadership	Increased diversity in parties' candidates and leadership	• Tenure and turnover of party leadership and candidates for office • Interviews and focus groups of party members about fairness of system
Programming Aimed at Institutions and Institutional Reform					
Campaign and party finance reform	Parties participating in dialogue	Parties support reform commission	• Proposal drafted • Proposal approved	• Parties adjust finance strategies • Parties report to electoral authority	• Published campaign finance reports • Interviews to determine changed

					finance strategies
Programming Aimed at Non-Partisan Actors					
Assist electoral authority to: develop and validate voter registration, oversee party and campaign finance, recruit and train poll watchers	Dialogues including parties and electoral authority	• Develop strategies for: ○ building voter registration lists and strategy for verification, including sources of funding for the effort ○ oversight of financial reporting, ○ pollwatcher training	• Implementation of system • resources dedicated • creation of new and credible lists of voters	• Increased voting participation • Reduced complaints of corrupt or partisan influences on registration system • Improved management of electoral process	• Voter participation rates • Formal complaints issued by voters or parties regarding registration
Organizational and Technical Capacity					
Programming Aimed at Partisan Actors					
Parties Develop Strategic Plans	Number of parties receiving training. Leadership positions of participants.	Participants outline strategic plan as part of training	Parties develop new strategic news	Parties revise strategies, in line with strategic plans	• Interviews about pre-training strategic plans • New strategic plan (if not confidential)
Parties develop candidate training programs	Number of parties participating	Participants outline parties' needs & learn about potential training systems	Parties develop program, allocate funds and personnel for it, and begin training	Candidates learn new campaign strategies	• Interviews regarding pre-training procedures • Interviews with trained and not-trained candidates
Parties improve accounting procedures	Number of parties receiving training	Parties identify weaknesses (and strengths) in their procedures, and gain understanding of sound principles.	• Parties revise accounting procedures	• Allocation of funds within party changes • Corruption declines	• Interviews about pre- and post-training procedures and changes in spending patterns

* Most of these data needs presupposed gathering data for treatment and non-treatment groups, perhaps comparing regions that did and did not receive party aid programs. The column provides examples, and is not intended to be comprehensive. Note also that most, but not all of the "special data needs" are pertinent to tests for impacts or higher level impacts, because data needs for outputs and outcomes is more self-evident

Appendix 2A: Worksheet for Performance Evaluation, by Development Characteristics (for Party System)

Development Characteristics	Goal Identified in RFA or Work Plan? (y/n + details)	Level of Priority/ Effort & Explanation	Activities designed to address goal	Suggested Data Needs and Indicator(s)
Accountability, Representation, and Participation				
Accountability				
Voters identify with parties, and demand services and responses to policy concerns				Surveys or focus groups testing voters' attachment to and knowledge of parties (beyond electoral choices). Questions also about contact with parties and about constituency service.
Transparency in government activities				• Publications of legislative roll-call votes. • Activities of ombudsman and legislative watchdog groups. • Interviews with experts.
Allow voters diverse choices among parties				• Number of parties, by region. • Electoral data showing geographic and ethnic support bases of parties.
Voters evaluate parties based on policy performance (retrospective voting)				Survey data (or focus groups)
Electoral system allows voters to hold leaders				Expert analysis of electoral system design

accountable				
Voters support democracy and suggest they trust the democratic processes				Survey data and focus groups
Representation				
Number of parties balances representation and governability				Effective and raw number of Parties (registered and competing), by region/district
Allow voters choice among candidates within parties				Analysis of candidate selection system and ballot system
Range of parties balances regional or ascriptive identification with nationally-oriented catch-all parties				Electoral and survey data capturing ethnic and regional voting patterns
Participation				
High level of voter knowledge and participation with limited polarization				• Surveys & Focus Groups testing voter knowledge of parties, politics, and government • Surveys asking voters' about their discussions about politics and participation in politics beyond voting
High citizen trust in government decisions				Survey and focus groups
Limited electoral volatility				• Electoral data; • Pedersen index of volatility
High participation of women and underrepresented groups in legislature				• Gender and ethnic composition of regional and national legislature • Similar data for committee assignments & leadership positions
Multiparty caucuses (eg of				• Review of legislative caucuses

women) form in the legislature or as civil society organizations to advance policies				and their activities. • Lists of politically-oriented NGOs.
Governability and Good Governance				
Coalitions form in the legislature to pursue policy objectives; includes proposals of legislation and engaging the executive in substantive policy debates				• Composition of cabinet; • Roll call votes & studies of budget or other important bills to indicating role of opposition in supporting (or denying) legislation • Success rate and importance of legislative bills
Legislature forms multipartisan-coalitions and other structures to oversee executive (or other party) actions				• Passage of transparency laws, freedom of information acts, and informal practices regarding these systems; • Questioning of ministers; • legislative role in exposing government corruption or decisions • Budgets for congressional oversight committees or bureaucratic offices
Legislature has mechanisms for bill analysis				• Review of technocratic support of legislature in committees or party offices (budgets for technocrats)
Stable and Peaceful Contestation				
Vibrant multiparty competition, but with limited electoral volatility				• Effective number of partied • Pedersen index, and change in support of major parties
High Voting Participation (but limited polarization)				• Electoral participation • Surveys indicating level of disagreement among voters

Increasing information to voters to counteract effects of volatility				Importance of incumbency (partisan or legislator) and local issues in comparison with national politics in electoral choices
Rule of Law and Free and Honest Elections				
Voters accept national electoral outcomes				Degree of protests after elections & party leader involvement in support of protests
Parties help to validate and then accept electoral results and support peaceful transition of power				• Change of partisan control of government • Interviews concerning previous transfers of power
Voters (correctly) perceive parties as adhering to rules of the game (limited corruption)				• Surveys and focus groups • Corruption indices • Expert interviews
There exists & parties support an impartial legal framework for elections and political parties				• Parties debate weaknesses in party or electoral law and propose improvements
There exists sustainable, indigenous capacity to effectively administer elections				• Expert interviews about election authority (national and regional)
Parties all support local, regional, and international efforts to monitor electoral processes				• Coverage of polling stations by individual parties; participation (and level) by NGOs and international organizations
Parties support electoral authorities attempts to: o Develop comprehensive and impartial voter lists o deploy non-partisan				• Parties' statements about voter lists • Numbers of non-partisan poll watchers • Explanation of dispute settlement system,

○ electoral observers run transparent electoral process				• Penalties (and enforcement) for breaking finance or other rules;
Parties support campaign finance system that provides them and their competitors a legal and transparent way to raise funds				• Rules of campaign finance • Interviews with parties, electoral authorities, and observers about common sources of funds, abuses, and oversight

Appendix 2B: Worksheet for Performance Evaluation, by Development Characteristics (Party Level)

Development Characteristics	Goal Identified in RFA or Work Plan? (y/n + details)	Level of Priority/ Effort (Low-High) & Explanation	Activities designed to address goal	Suggested Data Needs and Indicator(s)
Accountability, Representation, and Participation				
Accountability				
Parties base platforms and seek votes based on policy positions and policy effectiveness, not (only) identity				Survey data to evaluate extent of Retrospective Voting Platforms; interviews about policy focus of campaigns
Disseminate information about policy positions and activities to				Number, organization, and budgets of regional offices; data on contact with voters or other activities in those offices

constituents				
Open their decisionmaking processes				Surveys with mid-level party officials
Parties mobilize voters to press demands without systemic destabilization				Expert surveys
Parties respond to local-level and national citizen concerns				Party Nationalization; policy proposals to evaluate regional targets of programs
Parties develop transparent methods for selection of qualified candidates				Formal and informal rules for candidate choice. Competitiveness of primaries or conventions Background data on candidates and elected officials
Leaders are held accountable to membership via use of transparent selection methods				Electoral rules; leadership turnover in party and government posts; reelection rates of leaders vs rank & file
Party members cohere around ideology and policy positions (and its legislators thus vote in a unified manner), but are tolerant of dissent. Positions are not dictated by a leader.				Roll call (Rice) scores; interviews about leaders' powers to enforce discipline; elite surveys to assess degree of ideological or policy agreement among party's legislators
Representation				

Parties link national and regional constituencies				Party platforms and advertisements; expert analyses
Participation of historically excluded populations increases in internal party decisions				Parties' membership data, with detail on gender, youth, and minority groups. Composition (by these groups) of parties' internal committees, leadership posts, and candidates.
Participation				
Increase voters' identification with parties and demands for party services				Surveys testing voter ties to parties
Parties seek citizen input and respond to their concerns				Data on constituency service; interviews about forms of citizen contact
Inform and empower citizens to participate in the political process				Interviews about outreach efforts
Number of women, youth, and underrepresented groups increases as candidates and as party leaders				Data on participation in party-run training sessions for traditionally excluded groups
Parties respond to voter concerns for national level policy reform				Parties develop and disseminate positions and proposed legislation on national priorities
Voters identify with parties due to ideological positions and/or support of				Survey data to match voters' partisan preferences voters' ideology, and change in partisan preferences over time; electoral

community issues				data on split-ticket voting can also help to distinguish preferences for the party versus a particular candidate
Citizens and groups contact, request, or demand services from parties				Parties' records about these contacts; interviews with social groups
Parties develop national constituency without ignoring local representation duties (nationalization)				Regional level electoral data to evaluate party nationalization
Governability and Good Governance				
Parties develop capacity for bill analysis and developing legislation				Interviews about how parties analyze policy; experience of party members dedicated to this task
Parties negotiate for compromises (sacrificing ideology battles for policy accords)				Roll call or other data to show which parties' supported important legislation; co-signing of proposals; records on participation in policy forums
Stable Peaceful Contestation				
Parties run aggressive but "responsible" campaigns				Interview evidence on parties' campaign strategies
Parties develop roots in society				Interview evidence of parties' participation with local groups; presence of partisan activities during non-electoral seasons
Parties build support based on long track				Policy specificity in platforms; surveys to assess whether voters

record of policy				choose parties who are closely aligned with them on policy issues (policy vs identity voting)
Rule of Law and Free and Honest Elections				
Internal party elections are run fairly and constituents accept outcomes				Role of electoral authority in overseeing parties' internal elections; interviews with party members about competiveness of internal elections
Parties deploy trained poll watchers to all polling places				Records on number of partisan poll workers deployed; training programs for those poll watchers
Parties participate in national debates about electoral processes				Number of fora and interview data about substantive role of participants
Parties use exit polls to validate results				Interview or media reports regarding party statements about validity of electoral outcomes
Parties raise sufficient funds for campaigns and non-campaign activities through legal and transparent means				Survey data on perceptions of corruption in parties

Appendix 3: CEPPS Program Goals

Developing an impartial legal framework for elections and political parties
Building a sustainable, indigenous capacity to effectively administer elections
Informing and empowering all members of society to become active participants in the political process
Supporting local, regional and international efforts to monitor electoral processes
Ensuring inclusive political participation wherein a greater range of groups, including women and other historically excluded populations, influence and participate in political processes
Promoting consensus-building for peaceful agreement on democratic reform, rules and framework through peaceful, broad-based participation in defining and negotiating changes to governing structures
Strengthening political parties to ensure a competitive multi-party system that represents the diversity within a country
Fostering the smooth transfer of power following elections
Strengthening governance capacity of elected leaders and bodies
Building capable and sustainable local and regional civil society organizations engaged in election assistance and in providing technical assistance to political parties and political processes

Glossary for Evaluation Methodology

Control Group: is identical to all other items or subjects that a researcher examines in an experiment with the exception that it does not receive the treatment or the experimental manipulation that the treatment group receives. A control group is used as a baseline measure.

Confounding Factors: is a variable related to one or more of the variables defined in a study. It may mask an actual association or suggest a causal relationship between the independent and dependent variables where no real association between them exists. If confounding factors are not measured and considered, bias may result in the conclusion of the study.

Counterfactual: is a conditional statement that expresses what has not happened but could, would, or might under differing conditions. It is a claim contrary to the facts, as "if that president had not been elected, a civil war would not have started."

Difference-in-difference Design: is a quasi-experimental technique that measures the effect of a treatment at a given period in time. In contrast to a within-subjects estimate of the treatment effect (that measures the difference in an outcome after and before treatment) or a between-subjects estimate of the treatment effect (that measures the difference in an outcome between the treatment and control groups), this estimator represents the difference between the pre-post, within-subjects differences of the treatment and control groups.

Effective Number of Parties (ENP): is a method of counting the number of parties that takes account of the votes won or seats held by each party. The statistic does not presuppose a "best" type of system; it is simply an analytical tool that supports cross-time or cross-country comparisons. The ENP is calculated by squaring the vote percentage of each party, summing those values, and then taking the inverse. For example, if there are 3 parties which won 30, 20, and 50 percent of the vote, the formula is simply: $1/(.30^2+.20^2+.50^2) = 2.63$. (There are also alternative formulae.)

Encouragement Design: sometimes the subjects of a study need to be "encouraged" to participate with some kind of stimulus (e.g., financial) because it is not clear that randomly selected units to a treatment will participate, or it is not possible to prevent non-selected units from participating.

Fundamental Problem of Causal Inference: is the proposition that the true effect of any intervention is an intrinsically unobservable quantity.

Hawthorne Effect: it occurs when individuals alter their behavior as a result of their possible knowledge that they are being treated.

Higher-Level Impacts: it refers to indicators used in USAID's monitoring and evaluations efforts that measure the results of behavior or policy changes. For instance, they measure changes in party behavior (e.g., whether a new law leads to a reduction in the number of parties).

Instrumental Variables: is a method used to estimate causal relationships when controlled experiments are not feasible. When the explanatory variables are correlated with the error terms, ordinary linear regression generally produces biased and inconsistent estimates. In those cases, the use of an instrumental variable, which does not itself belong in the explanatory equation and is correlated with the endogenous explanatory variables, allows consistent estimation.

John Henry Effect: it occurs when individuals in the control group may change their behavior or try harder to achieve some outcome because they were passed over for treatment.

Oversubscription Design: it is used for conducting a randomizing control trial. The main advantages of this technique are that the program implementers can control who receives the treatment and can ensure that the treatment is administered through random assignment from the pool of potential participants.

Phased Roll-Out Design: when it is not possible to draw treatment and control groups simply at random, randomization may still be possible through this design, in which the order with which units receive a treatment is randomly determined.

Propensity Score Matching: is a method to correct for sample selection bias due to observable differences between the treatment and control groups. Randomized experiments enable unbiased estimation of treatment effects. However, for observational experiments, the assignment of units to the control and treatment groups is not randomized, but haphazard. Matching involves pairing treatment and control units that are similar in terms of their observable characteristics, which attempts to reduce the effect of confounding factors.

Quasi-Experimental Designs: involves selecting control and treatment groups, upon which a variable is tested, without any random pre-selection processes.

Random Assignment: Refers to the use of chance procedures in experiments to ensure that each participant or unit of analysis has the same opportunity to be assigned to any given group. When the unit is persons, participants are randomly assigned to different groups, such as the treatment or control group. Random assignment might involve such tactics as flipping a coin, drawing names out of a hat or assigning random numbers to participants.

Selection Bias: Also referred as "the selection effect," it is a statistical bias in which there is an error in choosing the individuals or groups to take part in a scientific study. The term most often refers to the distortion of a statistical analysis, resulting from the

method of collecting samples. If the selection bias is not taken into account then any conclusions drawn may be wrong.

Self-Selection Bias: arises in any situation in which individuals select themselves into a group, causing a biased sample with nonprobability sampling. It is commonly used to describe situations where the characteristics of the people which cause them to select themselves in the group create abnormal or undesirable conditions in the group.

Spill-Over Effects: it refers to interferences between the control and the treatment groups.

Treatment Group: In experiments, the treatments (such as participation in a training session) are give to one group, but not to the control group.

Unobservable Variables: They are unknown variables that may differentiate treated and control groups, which may be highly relevant to a group's propensity to experience a given treatment. For instance, in cases where program participation is the result of self-selection into a treatment group, unobservable variables related to motivation or personality characteristics may be especially relevant in explaining change.