USAID
FROM THE AMERICAN PEOPLE

NUMBER 19

1ST EDITION, 2010 DRAFT

PERFORMANCE MONITORING & EVALUATION

TIPS

RIGOROUS IMPACT EVALUATION

ABOUT TIPS

These TIPS provide practical advice and suggestions to USAID managers on issues related to performance monitoring and evaluation. This publication is a supplemental reference to the Automated Directive System (ADS) Chapter 203.

WHAT IS RIGOROUS IMPACT EVALUATION?

Rigorous impact evaluations are useful for determining the effects of USAID programs on outcomes. This type of evaluation allows managers to test development hypotheses by comparing changes in one or more specific outcomes to changes that occur in the absence of the program. Evaluators term this the *counterfactual*. Rigorous impact evaluations typically use comparison groups, composed of individuals or communities that do not participate in the program. The comparison group is examined in relation to the treatment group to determine the effects of the USAID program or project.

FIGURE 1. DEFINITIONS OF IMPACT EVALUATION

- An evaluation that looks at the impact of an intervention on final welfare outcomes, rather than only at project outputs, or a *process evaluation* which focuses on implementation.
- An evaluation carried out some time (five to ten years) after the intervention has been completed, to allow time for impact to appear.
- An evaluation considering all interventions within a given sector or geographical area.
- An evaluation concerned with establishing the counterfactual, i.e., the difference the project made (how indicators behaved with the project compared to how they would have been without it).

Impact evaluations may be defined in a number of ways (see Figure 1). For purposes of this TIPS, rigorous impact evaluation is defined by the evaluation design (quasi-experimental and experimental) rather than the topic being evaluated. These methods can be used to attribute change at any program or project outcome level, including Intermediate Results (IR), sub-IRs, and Assistance Objectives (AO).

Decisions about whether a rigorous impact evaluation would be appropriate and what type of rigorous impact evaluation to conduct are best made during the program or project design phase, since many types of rigorous impact evaluation can only be utilized if comparison groups are established and baseline data is collected before a program or project intervention begins.

WHY ARE RIGOROUS IMPACT EVALUATIONS IMPORTANT?

A rigorous impact evaluation enables managers to determine the extent to which a USAID program or project actually caused observed changes.

A Performance Management Plan (PMP) should contain all of the tools necessary to track key objectives (see also TIPS 7 Preparing a Performance Management Plan). However, comparing data from performance indicators against baseline values demonstrates only whether change has occurred, with very little information about what actually *caused* the observed change. USAID program managers can only say that the program is correlated with changes in outcome, but cannot confidently attribute that change to the program.

FIGURE 2. A WORD ABOUT WORDS

Many of the terms used in rigorous evaluations hint at the origin of these methods: medical and laboratory experimental research. The activities of a program or project are often called the *intervention* or the *independent variable*, and the outcome variables of interest are known as *dependent variables*. The target population is the group of all individuals (if the *unit of analysis* or *unit* is the individual) who share certain characteristics sought by the program, whether or not those individuals actually participate in the program. Those from the target population who actually participate are known as the *treatment group*, and the group used to measure what would have happened to the treatment group had they not participated in the program (the *counterfactual*) is known as a *control group* if they are selected randomly, as in an *experimental evaluation*, or, more generally, as a *comparison group* if they are selected by other means, as in a *quasi-experimental evaluation*.

There are normally a number of factors, outside of the program, that might influence an outcome. These are called *confounding factors*. Examples of confounding factors include programs run by other donors, natural events (e.g., rainfall, drought, earthquake, etc.), government policy changes, or even maturation (the natural changes that happen in an individual or community over time). Because of the potential contribution of these confounding factors, the program manager cannot claim with full certainty that the program caused the observed changes or results.

In some cases, the intervention causes all observed change. That is, the group receiving USAID assistance will have improved significantly while a similar, non-participating group will have stayed roughly the same. In other situations, the target group may have already been improving and the program helped to accelerate that positive change. Rigorous evaluations are designed to identify the effects of the program of interest even in these cases, where both the target group and non-participating groups may have both changed, only at different rates. By identifying the effects caused by a program, rigorous evaluations help USAID, implementing partners and key stakeholders learn which program or approaches are most effective, which is critical for effective development programming.

WHEN SHOULD THESE METHODS BE USED?

Rigorous impact evaluations can yield very strong evidence of program effects. Nevertheless, this method is not appropriate for all situations. Rigorous impact evaluations often involve extra costs for data collection and always require careful planning during program implementation. To determine whether a rigorous impact evaluation is appropriate,

potential cost should be weighed against the need for and usefulness of the information.

Rigorous impact evaluations answer evaluation questions concerning the causal effects of a program. However, other evaluation designs may be more appropriate for answering other types of evaluation questions. For example, the analysis of 'why' and 'how' observed changes, particularly unintended changes, were produced may be more effectively answered using other evaluation methods, including participatory evaluations or rapid appraisals. Similarly, there are situations when rigorous evaluations, which often use comparison groups, will not be advisable, or even possible. For example, assistance focusing on political parties can be difficult to evaluate using rigorous methods, as this type of assistance is typically offered to all parties, making the identification of a comparison group difficult or impossible. Other methods may be more appropriate and yield conclusions with sufficient credibility for programmatic decision-making.

While rigorous impact evaluations are sometimes used to examine the effects of only one program or project approach, rigorous impact evaluations are also extremely useful for answering questions about the effectiveness of alternative approaches for achieving a given result, e.g., which of several approaches for improving farm productivity, or for delivering legal services, are most effective.

Missions should consider using rigorous evaluations strategically to answer specific questions about the effectiveness of key approaches. When multiple rigorous evaluations are carried out across Missions on a similar topic or approach, the results can be used to identify approaches that can be generalized to other settings, leading to significant advances in programmatic knowledge. Rigorous methods are often useful when:

- Multiple approaches to achieving desired results have been suggested, and it is unclear which approach is the most effective or efficient;
- An approach is likely to be replicated if successful, and clear evidence of program effects are desired before scaling up;
- A program uses a large amount of resources or affects a large number of people; and
- In general, little is known about the effects of an important program or approach, as is often the case with new or innovative approaches.

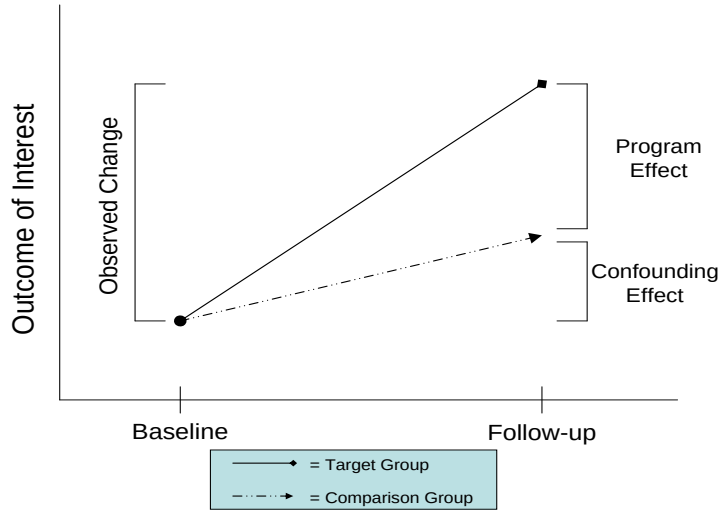
PLANNING

Rigorous methods require strong performance management systems to be built around a clear, logical results framework (see TIPS 13 Building a Results Framework). The development hypothesis should clearly define the logic of the program, with particular emphasis on the intervention (*independent variable*) and the principal anticipated results (*dependent variables*), and provides the basis for the questions that will be addressed by the rigorous evaluation.

Rigorous evaluation builds upon the indicators defined for each level of result, from inputs to outcomes, and requires high data quality. Because quasi-experimental and experimental designs typically answer very specific evaluation questions and are generally analyzed using quantitative methods, they can be paired with other evaluation tools and methods to provide context, triangulate evaluation conclusions, and examine how and why effects were produced (or not) by a program. This is termed mixed method evaluation (see TIPS 16, Mixed Method Evaluations).

Unlike most evaluations conducted by USAID, rigorous impact evaluations are usually only possible, and are always most effective, when planned before project implementation begins. Evaluators need time prior to implementation to identify appropriate indicators, identify a comparison group, and set baseline values. If rigorous evaluations are not planned prior to implementation, the number of potential evaluation design options is reduced, often leaving alternatives that are either more complicated or less rigorous. As a result, Missions should consider the feasibility of and need for a

FIGURE 3. CONFOUNDING EFFECTS



rigorous evaluation prior to and during project design.

DESIGN

Although there are many variations, rigorous evaluations are divided into two categories: *quasi-experimental* and *experimental*. Both categories of rigorous evaluations rely on the same basic concept - using the counterfactual to estimate the changes caused by the program. The counterfactual answers the question, *"What would have happened to program participants if they had not participated in the program?"* The comparison of the counterfactual to the observed change in the group receiving USAID assistance is the true measurement of a program's effects.

While before and after measurements of a single group using a baseline allow the measurement of a single group both with and without program participation, this design does not control for all the other confounding factors that might influence the participating group during program implementation. Well constructed, comparison groups provide a clear picture of the effects of program or project interventions on the target group by differentiating program/project effects from the effects of multiple other factors in the environment that affect both the target and comparison groups. This means that in situations where economic or other factors affecting both groups make everyone better off, it will still be possible to see the additional or incremental improvement caused by the program or project, as Figure 3 illustrates.

QUASI-EXPERIMENTAL EVALUATIONS

To estimate program effects, quasi-experimental designs rely on measurements of a non-randomly selected comparison group. The most common means for selecting a comparison group is *matching,* wherein the evaluator 'hand-picks' a group of similar units based on observable characteristics that are thought to influence the outcome. For example, the evaluation of an agriculture program aimed at increasing crop yield might seek to compare participating communities against other communities with similar weather patterns, soil types, and traditional crops, as communities sharing these critical characteristics would be most likely to behave similarly to the treatment group in the absence of the program.

However, program participants are often selected based on certain characteristics, whether it is level of need, motivation, location, social or political factors, or some other factor. While evaluators can often identify and match many of these variables, it is impossible to match all factors that might create differences between the treatment and comparison groups, particularly characteristics that are more difficult to measure or are *unobservable,* such as motivation or social cohesion. For example, if a program is targeted at

WHAT IS EXPERIMENTAL AND QUASI-EXPERIMENTAL EVALUATION?

Experimental design is based on a the selection of the comparison and treatment group through random sampling.

Quasi-experimental design is based on a comparison group that is chosen by the evaluator (that is, not based on random sampling).

FIGURE 4.
QUASI-EXPERIMENTAL EVALUATION OF THE KENYA NATIONAL CIVIC EDUCATION PROGRAM PHASE II (NCEP II)

NCEP II, funded by USAID in collaboration with other donors, reached an estimated 10 million individuals through workshops, drama events, cultural gatherings and mass media campaigns aimed at changing individuals' awareness, competence and engagement in issues related to democracy, human rights, governance, constitutionalism, and nation-building. To determine the program's impacts on these outcomes of interest, NCEP as evaluated using a *quasi-experimental* design with a matched comparison group.

Evaluators matched participants to a comparison group of non-participating individuals who shared geographic and demographic characteristics (such as age, gender, education, and involvement with CSOs). This comparison group was compared to the treatment group along the outcomes of interest to identify program effects. The evaluators found that the program had significant long term effects, particularly on 'civic competence and involvement' and 'identity and ethnic group relations, but had only negligible impact on 'Democratic Values, Rights, and Responsibilities'. The design also allowed the evaluators to assess the conditions under which the program was most successful. They found confirmation of prior assertions of the critical role in creating lasting impact of multiple exposures to civic education programs through multiple participatory methods.

- 'The Impact of the Second National Kenya Civic Education Programme (NECP II-URAIA) on Democratic Attitudes, Values, and Behavior', Steven E. Finkel and Jeremy Horowitz, MSI

communities that are likely succeed, then the target group might be expected to improve relative to a comparison group that was not chosen based on the same factors. Failing to account for this in the selection of the comparison group would lead to a biased estimate of program impact. ***Selection bias*** is the difference between the comparison group and the treatment group caused by the inability to completely match on all characteristics, and the uncertainty or error this generates in the measurement of program effects.

Other common quasi-experimental designs, in addition to matching, are described below.

Non-Equivalent Group Design. This is the most common quasi-experimental design in which a comparison group is hand-picked to match the treatment group as closely as possible. Since hand-picking the comparison group cannot completely match all characteristics with the treatment group, the groups are considered to be 'non-equivalent'.

Regression Discontinuity. Programs often have eligibility criteria based on a cut-off score or value of a targeting variable. Examples include programs accepting only households with income below 2,000 USD, organizations registered for at least two years, or applicants scoring above a 65 on a pre-test. In each of these cases, it is likely that individuals or organizations just above and just below the cut-off value would demonstrate only marginal or incremental differences in the absence of USAID assistance, as families earning 2,001 USD compared to 1,999 USD are unlikely to be significantly different *except* in terms of eligibility for the program. Because of this, the group just above the cut-off serves as a comparison group for those just below (or vice versa) in a regression discontinuity design.

Propensity Score Matching. This method is based on the same rationale as regular matching: a comparison group is selected based on shared observable characteristics with the treatment group. However, rather than 'hand-picking' matches based on a small number of variables, propensity score matching uses a statistical process to combine information from all data collected on the target population to create the most accurate matches possible based on observable characteristics.

Interrupted Time Series.[1] Some programs will encounter situations where a comparison group is not possible, often because the intervention affects everyone at once, as is typically the case with policy change. In these cases, data on the outcome of interest are recorded at numerous intervals before and after the program or activity take places. The data form a *time-series* or trend, which the evaluator analyzes for significant changes around the time of the intervention. Large spikes or drops immediately after the intervention signal changes caused by the program. This method is slightly different from the other rigorous methods as it does not use a comparison group to rule out potentially confounding factors, leading to increased uncertainty in evaluation conclusions. Interrupted time series are most effective when data are collected regularly both before and after the intervention, leading to a long time series, and alternative causes are monitored.

EXPERIMENTAL EVALUATION

In an experimental evaluation, the treatment and comparison groups are selected from the target population by a random process. For example, from a target population of 50 communities that meet the eligibility (or targeting) criteria of a program, the evaluator uses a coin flip, lottery, computer program, or some other random process *to determin*e the 25 communities that will participate in the program (*treatment group*) and the 25 communities that will not (*control group,* as the comparison group is called when it is selected randomly). Because they use random selection processes, experimental evaluations are often called *randomized evaluations* or *randomized controlled trials* (RCTs).

Random selection from a target population into treatment and control groups is the most effective tool for eliminating selection bias because it removes the possibility of any individual characteristic influencing selection. Because units are not assigned to treatment or control groups based on specific characteristics, but rather are divided randomly, all characteristics that might lead to selection bias, such as motivation, poverty level, or proximity, will be roughly equally divided between the treatment and control groups. If an evaluator uses random assignment to determine treatment and control groups, she might, by chance, get two or three very motivated communities in a row assigned to the treatment group, but if the program is working in more than a handful of communities, the number of motivated communities will likely balance out between treatment and control in the end.

Because random selection completely eliminates selection bias, experimental evaluations are often easier to analyze and provide more credible evidence than quasi experimental designs. Random assignment can be done with any type of unit, whether the unit is the individual, groups of individuals (e.g., communities or districts), organizations, or facilities (e.g., health center or school) and usually follows one of the designs discussed below.

Simple Random Assignment. When the number of program participants has been decided and additional eligible individuals are identified, simple random assignment through a coin flip or lottery can be used to select the treatment group and control groups. Programs often encounter 'excess demand' naturally (for example in training programs, participation in study tours, or where resources limit the number of partner organizations), and simple random assignment can be an easy and fair way to determine participation while maximizing the potential for credible evaluation conclusions.

Phased-In Selection. In some programs, the delivery of the intervention does not begin everywhere at the same time. For capacity or logistical reasons, some units receive the program intervention earlier than others. This type of schedule creates a natural opportunity for using an

[1] Interrupted time series is normally viewed as a type of impact evaluation. It is typically considered quasi-experiemental although it does not use a comparison group.

FIGURE 5.
EXPERIMENTAL EVALUATION OF THE IMPACTS OF EXPANDING CREDIT ACCESS IN SOUTH AFRICA

While commercial loans are a central component of most microfinance strategies, there is much less consensus on whether consumer loans are also for economic development. Microfinance in the form loans for household consumption or investment has been criticized as unproductive, usurious, and a contributor to debt cycles or traps. In an evaluation partially funded by USAID, researchers used an experimental evaluation designed to test the impacts of access to consumer loans on household consumption, investment, education, health, wealth, and well-being.

From a group of 787 applicants who were just below the credit score needed for loan acceptance, the researchers randomly selected 325 (treatment group) that would be approved for a loan. The treatment group was surveyed, along with the remaining 462 who were randomly denied (control group), eight months after their loan application to estimate the effects of receiving access to consumer credit. The evaluators found that the treatment group was more likely to retain wage employment, less likely to experience severe hunger in their households, and less likely to be impoverished than the control group providing strong evidence of the benefits of expanding access to consumer loans.

-'Expanding Credit Access: Estimating the Impacts', Dean Karlan and Jonathan Zinman, http://www.povertyactionlab.org/projects/print.php?pid=62

experimental design. Consider a project where the delivery of a radio-based civic education program was scheduled to operate in 100 communities during year one, another 100 during year two, and a final 100 during year three. The year of participation can be randomly assigned. Communities selected to participate in year one would be designated as the first treatment group (T_1). For that year, all the other communities that would participate in Years Two and Three form the initial control group. In the second year, the next 100 communities would become the second treatment group (T_2), while the final 100 communities would continue to serve as the control group. Random assignment to the year of participation ensures that all communities will participate in the program but also maximizes evaluation rigor by reducing selection bias, which could be significant if only the most motivated communities participate in Year One.

Blocked (or Stratified) Assignment. When it is known in advance that the units to which a program intervention could be delivered differ in one or more ways that might influence the program outcome, (e.g., age, size of the community in which they are located, ethnicity, etc.), evaluators may wish to take extra steps to ensure that such conditions are evenly distributed between an evaluation's treatment and control groups. In a simple block (stratified) design, an evaluation might separate men and women, and then use randomized assignment within each block to construct the evaluation's treatment and control groups, thus ensuring a specified number or percentage of men and women in each group.

Multiple Treatments. It is possible that multiple approaches will be proposed or implemented for the achievement of a given result. If a program is interested in testing the relative effectiveness of three different strategies or approaches, eligible units can be randomly divided into three groups. Each group participates in one approach, and the results can be compared to determine which approach is most effective. Variations on this design can include additional groups to test combined or holistic approaches and a control group to test the overall effectiveness of each approach.

COMMON QUESTIONS AND CHALLENGES

While rigorous evaluations require significant attention to detail in advance, they need not be impossibly complex. Many of the most common questions and challenges can be anticipated and minimized.

COST

Rigorous evaluations will almost always cost more than standard evaluations that do not require comparison groups. However, the additional cost can sometimes be quite low depending on the type and availability of data to be collected. Moreover, findings from rigorous evaluations may lead to future cost-savings, through improved programming and more efficient use of resources over the longer term. Nevertheless, program managers must anticipate these additional costs, including the additional planning requirements, in terms of staffing and budget needs.

ETHICS

The use of comparison groups is sometimes criticized for denying treatment to potential beneficiaries. However, every program has finite resources and must select a limited number of program participants. Random selection of program participants is often viewed, even by those beneficiaries who are not selected, as being the fairest and most transparent method for determining participation.

A second, more powerful, ethical question emerges when a program seeks to target participants that are thought to be most in need of the program. In some cases, rigorous evaluations require a relaxing of targeting requirements (as discussed in Figure 6) in order to identify enough similar units to constitute a comparison group, meaning that perhaps some of those identified as the 'neediest' might be assigned to the comparison group. However, it is often the case that the criteria used to target groups do not provide a degree of precision required to confidently rank-order potential participants. Moreover, rigorous evaluations can help identify which groups benefit most, thereby improving targeting for future programs.

SPILLOVER

Programs are often designed to incorporate 'multiplier effects' whereby program effects in one community naturally spread to others nearby. While these effects help to broaden the impact of a program, they can result in bias in conclusions when the effects on the treatment group *spillover* to the comparison group. When comparison groups also benefit from a program, then they no longer measure only the confounding effects, but also a portion of the program effect. This leads to underestimation of program impact since they appear better off than they would have been in the absence of the program. In some cases, spillovers can be mapped and measured but, most often, they must be controlled in advance by selecting treatment and control groups or units that are unlikely to significantly interact with one another. A special case of spillover occurs in *substitution bias* wherein governments or other donors target only the comparison group to fill in gaps of service. This is best avoided by ensuring coordination between

FIGURE 6. TARGETING IN RIGOROUS EVALUATIONS

Programs often have specific eligibility requirements without which a potential participant could not feasibly participate. Other programs target certain groups because of perceived need or likelihood of success. Targeting is still possible with rigorous evaluations, whether experimental or quasi-experimental, but must be approached in a slightly different manner. If a program intends to work in 25 communities, rather than defining one group of 25 communities that meet the criteria and participate in the program, it might be necessary to identify a group of 50 communities that meet the eligibility or targeting criteria and will be split into the treatment and comparison group. This reduces the potential for selection bias while still permitting the program to target certain groups. In situations where no additional communities meet the eligibility criteria and the criteria cannot be relaxed, phase-in or multiple treatment approaches, as discussed below, might be appropriate.

the program and other development actors.

SAMPLE SIZE

During the analysis phase, rigorous evaluations typically use statistical tests to determine whether any observed differences between treatment and comparison groups represent actual differences (that would then, in a well designed evaluation, be attributed to the program) or whether the difference could have occurred due to chance alone. The ability to make this distinction depends principally on the size of the change and the total number of units in the treatment and comparison groups, or *sample size*. The more units, or higher the sample size, the easier it is to attribute change to the program rather than to random variations. During the design phase, rigorous impact evaluations typically calculate the number of units (or sample size) required to confidently identify changes of the size anticipated by the program. An adequate sample size helps prevent declaring a successful project ineffectual (false negative) or declaring an ineffectual project successful (false positive). Although sample size calculations should be done before each program, as a rule of thumb, rigorous impact evaluations are rarely undertaken with less than 50 units of analysis.

RESOURCES

This TIPS is intended to provide an introduction to rigorous impact evaluations. Additional resources are provided on the next page for further reference.

Further Reference

Initiatives and Case Studies:

- Office of Management and Budget (OMB):
  - http://www.whitehouse.gov/OMB/part/2004_program_eval.pdf
  - http://www.whitehouse.gov/omb/assets/memoranda_2010/m10-01.pdf
- U.S. Government Accountability Office (GAO):
  - http://www.gao.gov/new.items/d1030.pdf
- USAID:
  - Evaluating Democracy and Governance Effectiveness (EDGE): http://www.usaid.gov/our_work/democracy_and_governance/technical_areas/dg_office/evaluation.html
  - Measure Evaluation: http://www.cpc.unc.edu/measure/approaches/evaluation/evaluation.html
  - The Private Sector Development (PSD) Impact Evaluation Initiative: www.microlinks.org/psdimpact

- Millennium Challenge Corporation (MCC) Impact Evaluations: http://www.mcc.gov/mcc/panda/activities/impactevaluation/index.shtml
- World Bank:
  - The Spanish Trust Fund for Impact Evaluation: http://web.worldbank.org/WBSITE/EXTERNAL/EXTABOUTUS/ORGANIZATION/EXTHDNETWORK/EXTHDOFFICE/0,,contentMDK:22383030~menuPK:6508083~pagePK:64168445~piPK:64168309~theSitePK:5485727,00.html
  - The Network of Networks on Impact Evaluation: http://www.worldbank.org/ieg/nonie/
  - The Development Impact Evaluation Initiative: http://web.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTDEVIMPEVAINI/0,,menuPK:3998281~pagePK:64168427~piPK:64168435~theSitePK:3998212,00.html
- Others:
  - Center for Global Development's 'Evaluation Gap Working Group': http://www.cgdev.org/section/initiatives/_active/evalgap
  - International Initiative for Impact Evaluation: http://www.3ieimpact.org/

Additional Information:

- Sample Size and Power Calculations:
  - http://www.statsoft.com/textbook/stpowan.html
  - http://www.mdrc.org/publications/437/full.pdf
- World Bank: 'Evaluating the Impact of Development Projects on Poverty: A Handbook for Practitioners':
  - http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTPOVERTY/EXTISPMA/0,,contentMDK:20194198~pagePK:148956~piPK:216618~theSitePK:384329,00.html

Poverty Action Lab's 'Evaluating Social Programs' Course: http://www.povertyactionlab.org/course/

For more information:
TIPS publications are available online at [insert website]

Acknowledgements:

Our thanks to those whose experience and insights helped shape this publication including USAID's Office of Management Policy, Budget and Performance (MPBP). This publication was written by Michael Duthie of Management Systems International.

Comments regarding this publication can be directed to:
Gerald Britan, Ph.D.
Tel: (202) 712-1158
gbritan@usaid.gov

Contracted under RAN-M-00-04-00049-A-FY0S-84
Integrated Managing for Results II