USAID
FROM THE AMERICAN PEOPLE

VERSION 1.0 | SEPTEMBER 2013

TECHNICAL NOTE

Impact Evaluations

Monitoring and Evaluation Series

This Note defines impact evaluations and discusses design and key planning considerations.

Technical Notes are published by the Bureau for Policy, Planning and Learning and provide key concepts and approaches to USAID staff and partners related to the Program Cycle. This Technical Note supplements USAID ADS Chapter 203 and replaces TIPS 18, Rigorous Evaluations.

INTRODUCTION

This Note defines impact evaluations, explains when they should be commissioned according to USAID policy and describes different designs for quasi-experimental and experimental impact evaluations. The USAID Automated Directives System (ADS) 203 defines **impact evaluations** as *those that measure the change in a development outcome that is attributable to a defined intervention. Impact evaluations are based on models of cause and effect and require a credible and rigorously defined counterfactual to control for factors other than the intervention that might account for the observed change.*

Decisions about whether an impact evaluation would be appropriate, and what type of impact evaluation to conduct, are best made early during the project design phase. Some impact evaluation designs can only be implemented if comparison groups are established and baseline data is collected before an intervention begins. Although they are most effective and sometimes only possible when planned before program implementation, impact evaluations can sometimes be used to measure changes that occur either during or after program implementation. In most cases, an expert should be consulted in advance to determine whether an impact evaluation will be feasible.

This note outlines key considerations that USAID staff and evaluators should take into account when planning for and designing impact evaluations. Those commissioning an evaluation should include the evaluator when making decisions about an intervention's targeting and implementation, and consider issues related to logistics, time and cost. Therefore although impact evaluations are a powerful tool to answer key questions about a particular intervention, they should be used selectively and only when appropriate in terms of purpose and funding.

WHAT IS AN IMPACT EVALUATION?

Impact evaluations are useful for determining the effect of USAID activities on specific outcomes of interest. They test USAID development hypotheses by comparing changes in one or more specific outcomes to what would have happened in the absence of the intervention, called the *counterfactual*. Impact evaluations use a *comparison group*, composed of individuals or communities where an intervention will not be implemented, and one or more *treatment groups*, composed of project beneficiaries or communities where an intervention is implemented. The comparison between the outcomes of interest in the treatment and comparison group creates the basis for determining the impact of the USAID intervention. An impact evaluation helps demonstrate *attribution* to the specific intervention by showing what would have occurred in its absence.

Most interventions track changes in key outcomes through performance monitoring, but comparing data from performance indicators against baseline values demonstrates only whether change has occurred, but does not establish what actually caused the observed change. *Confounding factors* include interventions run by other donors, natural events (e.g. rainfall, drought, earthquake, etc.), government policy changes, or natural changes that happen in an individual or community over time. Due to the potential effects of confounding factors, USAID managers cannot claim that their interventions actually caused the observed changes or results. In some cases, the intervention does cause all observed change. In these cases, 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 intervention helped to accelerate that positive change. Or, intended outcomes may appear to be negative (for instance, during an economic downturn), but comparison groups fare even worse. Impact evaluations are designed to identify the effects of the intervention of interest in all of these cases, where both the target group and non-participating groups may have changed, but at different rates. By identifying the effects caused by an intervention, impact evaluations help USAID, implementing partners, and key stakeholders learn which approaches are most effective. This is critical for determining future development programming and resource allocation.

QUESTIONS FROM USAID-FUNDED IMPACT EVALUATIONS

- What is the added value of the use of sports in workforce development programs for at-risk youth in Honduras and Guatemala? To what extent are program effects stronger or weaker for female, higher risk, younger, or less educated participants?
- To what extent were neighbors of beneficiaries positively or negatively affected by a livelihoods program in Ethiopia?
- Does training traditional leaders on human rights and peaceful conflict mitigation result in improvements in community leadership and dispute resolution? To what extent do top-down, horizontal, or bottom-up social pressures change the behavior of local leaders?

Note that the term "impact evaluation" involves a specialized meaning of the word “impact.” In common usage, “impact” could refer to high level results or long-term outcomes from an intervention. However, "impact evaluation" implies a structured test of one or more hypotheses underlying an intervention. Impact evaluations are characterized by a specific evaluation design (quasi-experimental or experimental) in order to answer a cause-and-effect question. These methods can be used to attribute change at any program or project outcome level, but typically focus on one specific activity. Impact evaluations typically collect and analyze quantitative data, but should also be informed by qualitative data collection methods as long as they are used to gather information from both treatment and comparison groups.

WHEN SHOULD IMPACT EVALUATIONS BE USED?

Impact evaluations *answer cause-and-effect questions* about intervention effects. While impact evaluations are sometimes used to examine the effects of only one intervention or project approach, they 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 impact evaluations strategically to answer specific questions about the effectiveness of key approaches. *External validity* - the extent to which evaluation results can be generalized to other settings, such as when an intervention is scaled up or attempted in other regions - is an important consideration for impact evaluations. Ways to ensure external validity include carrying out multiple impact evaluations across Missions on a similar topic or approach and making sure that the evaluation measures the effects of an intervention on different types of beneficiaries (across gender, age, socioeconomic groups, or other relevant factors). It is important for Missions to consult sector experts and coordinate with their Regional and Pillar Bureaus to ensure that they are contributing to a Bureau-wide learning and evaluation strategy.

Impact evaluations require strong performance monitoring systems to be built around a clear logical framework. The development hypothesis should clearly define the logic of the project, with particular emphasis on the intervention (independent variable) and the principle anticipated results (dependent variables), and provides the basis for the questions that will be addressed by the impact evaluation.

WHEN TO CONDUCT IEs

ADS 203 states that "any activity within a project involving *untested hypotheses* or demonstrating *new approaches that are anticipated to be expanded* in scale or scope through US Government foreign assistance or other funding sources will, if feasible, undergo an impact evaluation… Any activity or project designated as a 'pilot' or 'proof of concept' will fall under this requirement."

The World Bank has published the following guidelines for when an impact evaluation is appropriate:

- Is the intervention *INNOVATIVE*? Is it testing a new, promising approach?
- Is the intervention *REPLICABLE*? Can it be scaled up or can it be applied to a different setting?
- Is the intervention *STRATEGICALLY RELEVANT*? Is it a flagship intervention that requires substantial resources; covers, or could be expanded to cover, a large number of people; or could generate substantial savings?
- Is the intervention *UNTESTED*? That is, is very little known about the effectiveness of the intervention globally or in the specific context in which it is implemented?
- Is the intervention *INFLUENTIAL*? Will results be used to inform key policy decisions?

(Impact Evaluation in Practice, p. 11)

Impact evaluations are always most effective when planned before implementation begins. Evaluators need time prior to implementation to identify appropriate indicators, identify a comparison group, and set baseline values. In most cases they must coordinate the selection of a treatment and comparison group with the implementing partners. If impact 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 an impact evaluation prior to and during project design. On the other hand, interventions should not be evaluated too early in their "start-up phase," when the implementation details of the intervention are still being worked out. A good way to account for startup issues is to conduct a small pilot in a few communities (not included in the evaluation) before working with and conducting an evaluation of the full sample.

While impact evaluations do require advanced planning and significant attention to detail, they need not be impossibly complex, particularly since many of the most common questions and challenges can be anticipated and minimized with advanced planning. In all cases, USAID staff must coordinate between the evaluator, the implementer, and other stakeholders to identify an appropriate comparison or control group.

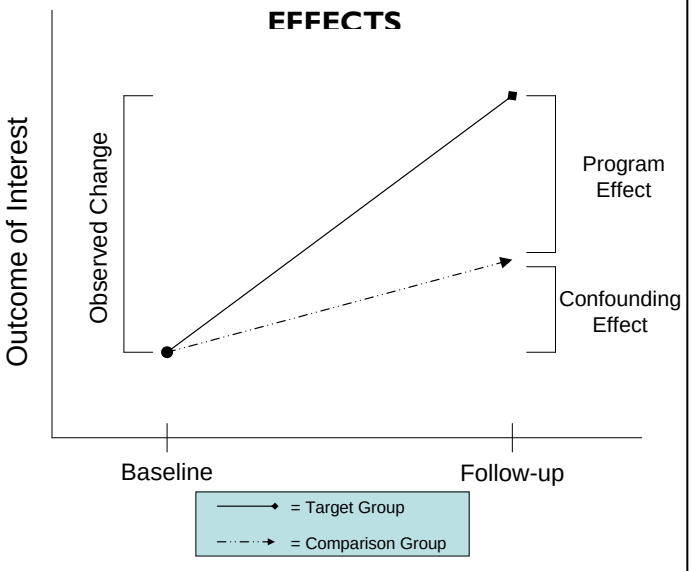
Finally, impact evaluations are not appropriate for all situations. They often involve extra costs for data collection and always require high levels of attention to detail, coordination, and time during intervention implementation. The potential extra costs should be considered against the information needs when determining whether an impact evaluation is appropriate. Performance evaluation may be more appropriate for answering other types of evaluation questions. For example, a USAID manager may be more interested in describing a process or analyzing 'why' and 'how' observed changes, particularly unintended changes, were produced. Questions generated in these cases may be more effectively answered using other evaluation methods, including participatory evaluations or rapid appraisals. Similarly, there are situations when impact evaluations, which use comparison or control groups, will not be advisable or even possible. For example, assistance focusing on political parties can be difficult to evaluate using impact evaluations, 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. Finally, when an intervention is offered in different ways across different sites (for example if communities select from a "package" of interventions) or changes significantly over time (for instance, when implementation details change significantly during the "start-up" phase of an activity), information from an impact evaluation will be less likely to apply to other settings or be useful in decisions about scale up.

DESIGN

This section outlines types of IE designs to increase understanding of what these approaches entail. Agency staff are encouraged to seek outside assistance from experts with evaluation methods training.

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

FIGURE 1: CONFOUNDING EFFECTS



Impact evaluations compare outcomes for groups that do and do not receive the intervention to answer questions about the counterfactual situation. While 'before-after' measurements of a single group using a baseline allow the measurement of a single group both with and without participation, this design does not control for all the other confounding factors that might influence the participating group during implementation. When well-constructed, comparison groups provide a clear picture of the effects of interventions on the target group by differentiating project effects from the effects of multiple other factors in

the environment which affect both the target and comparison groups. This means that in situations where economic or other factors that affect both groups are making everyone better off, it will still be possible to see the additional or incremental improvement caused by the intervention, as Figure 1 illustrates.

When a comparison group is generated using a random process, the evaluation is considered an *experimental evaluation* and the comparison group is referred to as a *control group*. When a comparison group is generated using other, non-random methods, the evaluation is considered a *quasi-experimental evaluation.*

QUASI-EXPERIMENTAL EVALUATIONS

To estimate intervention effects, quasi-experimental designs estimate the counterfactual by conducting measurements of a non-randomly selected comparison group. In many cases, intervention participants are selected based on certain characteristics, whether it is level of need, location, social or political factors, or some other factor. While evaluators can often identify and match many of these variables (or account for them in a regression analysis), it is impossible to match all factors that might create differences between the treatment and comparison groups, particularly characteristics which are more difficult to measure or are unobservable, such as motivation or social cohesion. For example, if an intervention is targeted at communities which are likely to succeed, then the target group might be expected to improve relative to a comparison group that was not chosen based on the same factors. On the other hand, if an intervention is targeted at the "neediest" potential beneficiaries, then the changes that the intervention expects to achieve may occur at a slower rate than with other, better-off individuals. Failing to account for this in the selection of the comparison group would lead to a biased estimate of intervention impact. *Selection bias* is the difference between the comparison group and the treatment group caused by the inherent differences between the two groups, and the uncertainty or error this generates in the measurement of intervention effects. All quasi-experimental evaluation designs must account for the extent to which they have minimized or measured selection bias.

Common quasi-experimental designs include:

- ***Non-Equivalent Group Design.*** In this design, 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'.

- ***Matching:*** The most common means for selecting a comparison group is matching, wherein the evaluator picks a group of similar units based on observable characteristics that are thought to influence the outcome. For example, the evaluation of an agriculture intervention 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 intervention. A type of matching design occurs when a comparison group is selected based on shared observable characteristics with the treatment group. However, rather than choosing 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. Neither type of matching can account for unobservable characteristics such as motivation.

- ***Regression Discontinuity.*** Interventions often have eligibility criteria based on a cut-off score or value of a targeting variable. Examples include interventions accepting only households with income below $2,000 USD, organizations or 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 intervention. 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.

In all of the above cases, the evaluation team should compare the treatment and comparison groups at ***baseline*** to make sure that the groups are in fact comparable. If there are significant differences at baseline in variables that may influence the outcome (for instance, the treatment group consists of wealthier communities) then the evaluation's ability to attribute later differences between the treatment and comparison groups to the intervention being evaluated will be less credible. If the evaluation is commissioned after the intervention begins, but baseline data is available, it is possible to conduct a *retrospective quasi-experimental design*.

EXAMPLE OF A QUASI-EXPERIMENTAL EVALUATION

USAID commissioned an impact evaluation of the Colombia Strategic Development Initiative, which provides U.S. assistance to the Colombian government's program to expand state presence in vulnerable areas and "consolidate" the rule of law. There are two separate mechanisms for this evaluation: A consortium of academics based at Princeton University and funded by the Department of Defense collaborated with USAID/Colombia's M&E program and gave technical advice to the firm that was contracted to conduct the evaluation.

The evaluation team used propensity score matching to identify municipalities that were similar to those selected for the program. They estimated propensity to receive treatment based on the historical presence of armed groups, market integration (or lack of), trends in contestation, presence of illicit crops, and population importance. They also measured trends in key outcome variables (from 2002 to 2010) to ensure that the treatment and comparison communities were in fact comparable. Data collection was conducted at the household (19,000), community, project, and municipal levels. The evaluation team had developed survey questions that had never been used before, in particular those that addressed sensitive issues such as participation or contact with armed guerrilla groups, so they piloted the questionnaire in one municipality before applying it to the entire evaluation sample.

The impact evaluation will allow both the Government of Colombia and USAID to learn which programs work where, and why. Given the substantive importance of the issue, as well as the resources invested in the programs by the USG globally, this is crucial. The fieldwork for the baseline was finished on June 2013 and two more waves of data collection are expected.

- ***Interrupted Time Series.*** In some situations, 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 is recorded at numerous intervals before and after the intervention takes place. 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 intervention. This method can be strengthened by the use of a comparison group to rule out potentially confounding factors, reducing uncertainty in evaluation conclusions. Interrupted time series are most effective when data is collected regularly both before and after the intervention, leading to a long time series, and when the analysis can account for alternative causes.

EXPERIMENTAL EVALUATION

In an experimental evaluation, the treatment and comparison groups are selected from the target population by a random process. Because the selection of treatment and control groups involves a random process, 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 2 or 3 very motivated communities in a row assigned to the treatment group, but if the intervention is working in more than a handful of communities, the number of motivated communities will likely balance out between treatment and control groups.

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 following designs:

WHAT UNIT TO RANDOMIZE?

A good rule of thumb is to randomize at the level in which the intervention takes place. For example, in an evaluation of a teacher training intervention, it would be impractical to ask the teacher to apply her new skills with some students and not others, and even if she could, selected students could influence their classmates anyway (see "spillover" below). Furthermore, it might not be politically or logistically feasible to train some teachers and not others within an individual school. It is more realistic to assign some schools to the treatment group and others to the control group, as long as the sample of schools is sufficiently large to detect statistically significant results. This type of decision is usually made in consultation with the evaluator, the implementing partner, and relevant USAID staff.

- ***Simple Random Assignment.*** When the number of intervention 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. Interventions often encounter or can generate 'excess demand' naturally, for example in training interventions, 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. For example, in a recently released USAID-funded impact evaluation conducted by the National Democratic Institute of a governance project in Cambodia, each field officer had to choose two communities that he or she felt that the project should work with. The evaluation team then flipped a coin for each pair, generating one treatment and one control community for each officer.

- ***Phased-In Selection.*** Even if an intervention plans to treat all eligible beneficiaries, there may be logistical reasons that prevent implementation from beginning everywhere at the same time. This type of schedule creates a natural opportunity for using an experimental design. Consider an intervention where delivery of a conditional cash transfer was scheduled to operate in 100 communities during year one, another 100 the second year and a final 100 in the intervention's third year. The year of participation can be randomly assigned. Communities selected to participate in Year 1 would be designated as the first treatment group (T1). For that year all the other communities, which 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 (T2), while the final 100 communities would continue to serve as the control group until the third year. This design is also known as *pipeline* or *stepped wedge design.*

- ***Randomized Promotion (Encouragement Design).*** In cases where randomized assignment is difficult, evaluators can randomize promotion of a particular intervention. For instance, a microfinance institution might be unwilling to turn potential clients away just because they are assigned to a control group, preferring to serve anyone who seeks to open a savings account. Evaluations of savings interventions instead randomly select some people within a community to receive a special invitation or incentive to open an account. If there is a substantial difference in uptake between those who receive an invitation to join and those who do not, then evaluators can compare the "invitation" and "no invitation" groups using an instrumental variable analysis (see above).

EXAMPLE OF A USAID EXPERIMENTAL EVALUATION

The evaluation of the A Ganar program is currently examining the impact of a sports-based youth workforce development program in Guatemala and Honduras on outcomes such as employment, school enrollment, and prevalence of risky behavior. The evaluation team piloted the study logistics and data collection with a small group (174 survey respondents in Honduras and Guatemala), which allowed them to: 1) refine the baseline survey and interview protocols; 2) determine the randomization strategy; and 3) work out the division of responsibilities with implementers. Local partners were willing to recruit a larger number of potential beneficiaries and allow the evaluation team to randomly allocate spots, but were not willing to exclude youth who had worked to recruit their peers into the program. Therefore they were allowed to select up to three youth to participate in the program – that is, not subject to randomized selection, and therefore included in the intervention but not in the evaluation sample. The rest were randomly assigned to treatment and control groups, and a subset of these was invited to participate in the baseline.

The full roll-out of the program will have a sample size of 1300 youth in Honduras, divided into one treatment and one control group (to test the effect of the program overall). In Guatemala, 1500 applicants will be divided into two treatment groups, one receiving a sports-based program and one receiving an equivalent program, and one control group. The design in Guatemala will allow the evaluation team to isolate the effect of the use of sports. Aside from the baseline survey, they will conduct two additional data collection events (immediate and long-term follow up) as well as qualitative longitudinal case studies to supplement quantitative findings.

- ***Blocked (or Stratified) Assignment.*** When it is known in advance that the units to which an intervention could be delivered differ in one or more ways that might influence the 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 often the case that multiple approaches will be proposed or implemented for the achievement of a given result. If an evaluation is interested in testing the relative effectiveness of 3 different strategies or approaches, eligible units can be randomly divided into 3 groups. An HIV prevention service, for example, could provide just prevention education to one group and prevention

education and peer support to another. Each group participates in one approach, and the results can be compared to determine which approach was 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. The multiple treatment groups can be generated using any of the methods outlined above.

ANALYSIS

In an impact evaluation, quantitative analysis can be as simple as comparing outcome means between treatment and comparison or control groups. When baseline measures exist, evaluators typically measure changes between baseline outcome measures and final outcome measures and compare these changes between treatment and control or comparison groups. This method allows them to take into account differences between the two groups that are constant over time and is known as a ***difference-in-difference*** analysis. Other analysis tools, such as ***multivariate regressions***, or *analyses of covariance* **(*ANCOVA*),** are more complex. Agency SOWs should require that evaluators report the results of analyses conducted using various tools and to use results from qualitative data collection to deepen explanations of findings.

KEY CONSIDERATIONS

EFFECT SIZE

In planning for an impact evaluation, it is important to clarify how large or small an *effect size* – that is, the magnitude of difference between the treatment and control group - the evaluator will be expected to measure. In theory, with unlimited evaluation funding and sample sizes, an impact evaluation could find that participants in the treatment group had a 0.001% higher income, but from a practical perspective, it is not worth determining whether an intervention has such a tiny impact. Considerations of effect size usually take into account what other interventions have accomplished given a certain level of funding and what has typically been achieved in a particular sector.

COST

Impact evaluations will almost always cost more than performance 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 impact evaluations may lead to future cost-savings, outweighing initial costs, through improved programming and more efficient use of resources. Nevertheless, USAID managers must anticipate these additional costs, including the additional staff resources implied by the level of attention to detail required, when considering and budgeting an impact evaluation. The largest cost of an impact evaluation is usually data collection, which in turn depends on the sample size (see below). PPL will provide additional guidance on budgeting for impact evaluations.

ETHICS

The use of comparison groups is often criticized for denying services to potential beneficiaries. This is less of a concern if the intervention has not been tested before, as there is also an ethical argument for demonstrating that an intervention does not have negative effects before implementing it at a widespread level. In addition, interventions can often take advantage of existing operational restrictions. For instance, most interventions have finite resources and must select a limited number of participants or geographic areas among those who would be eligible. In other cases, there is enough funding to work in an entire country but the implementer may not have the capacity to operate in all areas at once, which presents an opportunity to use a phased-in design. Random selection of participants or communities is often viewed, even by those beneficiaries who are not selected, as being the fairest and most transparent method for determining participation.

A second ethical question emerges when an intervention seeks to target participants that are thought to be

most in need. In some cases, impact evaluations require a relaxing of targeting requirements 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 are not definitively known and rarely with the degree of precision required to confidently rank-order potential participants. Alternatively, situations where the cutoff point for participation is such that those just below and just above are very similar to each other present an appropriate opportunity to use a regression discontinuity design.

Some countries require in-country ethical clearance for research. See *Protection of Human Subjects in Research Supported by USAID - A Mandatory Reference for ADS Chapter 200* for more information on protection of human subjects required by USAID. In cases where an evaluation firm hires an academic, they are usually required to secure clearance from their university's Human Subjects review board and provide evidence of having completed a human subjects training course.

SAMPLE SIZE

During the analysis phase, impact evaluations use statistical tests to determine whether any observed differences between treatment and comparison groups represent actual (statistically significant) differences 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. That is, there is always a chance that a group of communities that is randomly allocated to the treatment group may be more or less motivated (or more urban, or have another characteristic that influences the outcome) than the control group. With larger samples, this likelihood is reduced. The more units, or higher the sample size, the easier it is to attribute change to the intervention rather than to random variations. During the design phase, impact evaluations calculate the number of units (or sample size) required to confidently identify changes of the size anticipated by the intervention. An adequate sample size helps prevent declaring a successful intervention ineffectual (false negative) or declaring an ineffectual intervention successful (false positive). Sample size calculations should be done before each evaluation, in consultation with an expert, and take into account expected effect size and existing variability in a population. As a rule of thumb, impact evaluations are rarely undertaken with less than 100 (total) units of analysis.

SPILLOVER

Interventions are often designed to incorporate 'multiplier effects' whereby effects in one community naturally spread to others nearby. These effects help to broaden the impact of an intervention (and are desirable if the impact is positive), but they can result in bias in impact evaluation conclusions when the effects on the treatment group spillover to the comparison group. When comparison groups also benefit from an intervention, for example, this can lead to an underestimation of impact since they appear better off than they would have been in the absence of the intervention. 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. For example, it is usually more appropriate to divide classrooms or schools into treatment and control groups rather than individual students.

A special case of spillover occurs in substitution bias wherein governments or other organizations target only the comparison group to provide services similar to those provided to the treatment group(s). This is best avoided by ensuring coordination between USAID projects and other development actors in the region.

DISSEMINATION

Evaluations are only useful to the extent that results are available to interested stakeholders and decision makers. The ADS requires that evaluation results be posted to the DEC within 90 days of completion. Impact

evaluation contracts could also specify additional ways of disseminating results including: publications in academic journals, two-to-four page "policy briefs" with key findings, as well as conferences, workshops, and videos or other media. In some cases "pilot" projects funded by USAID may be scaled up by host country governments or other partners. In those cases, involving implementers and government stakeholders in the evaluation early on can ensure that they are invested in results and will increase their willingness to scale up successful projects.

ADDITIONAL RESOURCES

The following resources provide more information on impact evaluations.

Millennium Challenge Corporation (MCC)
- Evaluations Resource Page: http://www.mcc.gov/pages/results/evaluations

World Bank:
- World Bank Evaluation resources: http://go.worldbank.org/X5X013RJZ0
- Impact Evaluation in Practice: http://elibrary.worldbank.org/content/book/9780821385418
- Handbook for Impact Evaluation: http://go.worldbank.org/9H20R7VMP0
- The Strategic Trust Fund for Impact Evaluation: http://go.worldbank.org/Q2XYY39FW0
- The Development Impact Evaluation Initiative: http://go.worldbank.org/1F1W42VYV0

Abdul Latif Jameel Poverty Action Lab (JPAL)
- Methodology Resources: http://povertyactionlab.org/methodology
- 'Evaluating Social Programs' Course: http://www.povertyactionlab.org/course/ (An online version is available for free on I-tunes: https://itunes.apple.com/us/course/abdul-latif-jameel-poverty/id495065985)

InterAction
- Impact Evaluation Guidance Note and Seminar Series: http://www.interaction.org/impact-evaluation-notes

International Initiative for Impact Evaluation: http://www.3ieimpact.org/en/
- Theory-Based Impact Evaluation: Principles and Practice: http://www.3ieimpact.org/media/filer/2012/05/07/Working_Paper_3.pdf

Center for Global Development:
- 'Evaluation Gap Working Group': http://www.cgdev.org/section/initiatives/_active/evalgap
- When Does Rigorous Impact Evaluation Make a Difference? The Case of the Millennium Villages: http://www.cgdev.org/sites/default/files/1424496_file_Clemens_Demombynes_Evaluation_FINAL.pdf

USAID:
- Evaluation for Evaluation Specialists (EES) Course – http://university.usaid.gov/
- Value chain wiki: http://microlinks.kdid.org/good-practice-center/value-chain-wiki/impact-assessment
- E3 Trade Facilitation Monitoring, Evaluation and Learning Toolkit: http://usaidsite.carana.com/content/evaluation-pathway-4-rigorous-impact-evaluations
- Feed the Future M&E Guidance Series Volume 4 – Impact Evaluation http://www.feedthefuture.gov/sites/default/files/resource/files/Volume4_FTFImpact.pdf

Additional Information:
- Sample Size and Power Calculations:
- http://www.statsoft.com/textbook/stpowan.html
- http://www.mdrc.org/publication/core-analytics-randomized-experiments-social-research

ANNEX 1: DECISION AIDS

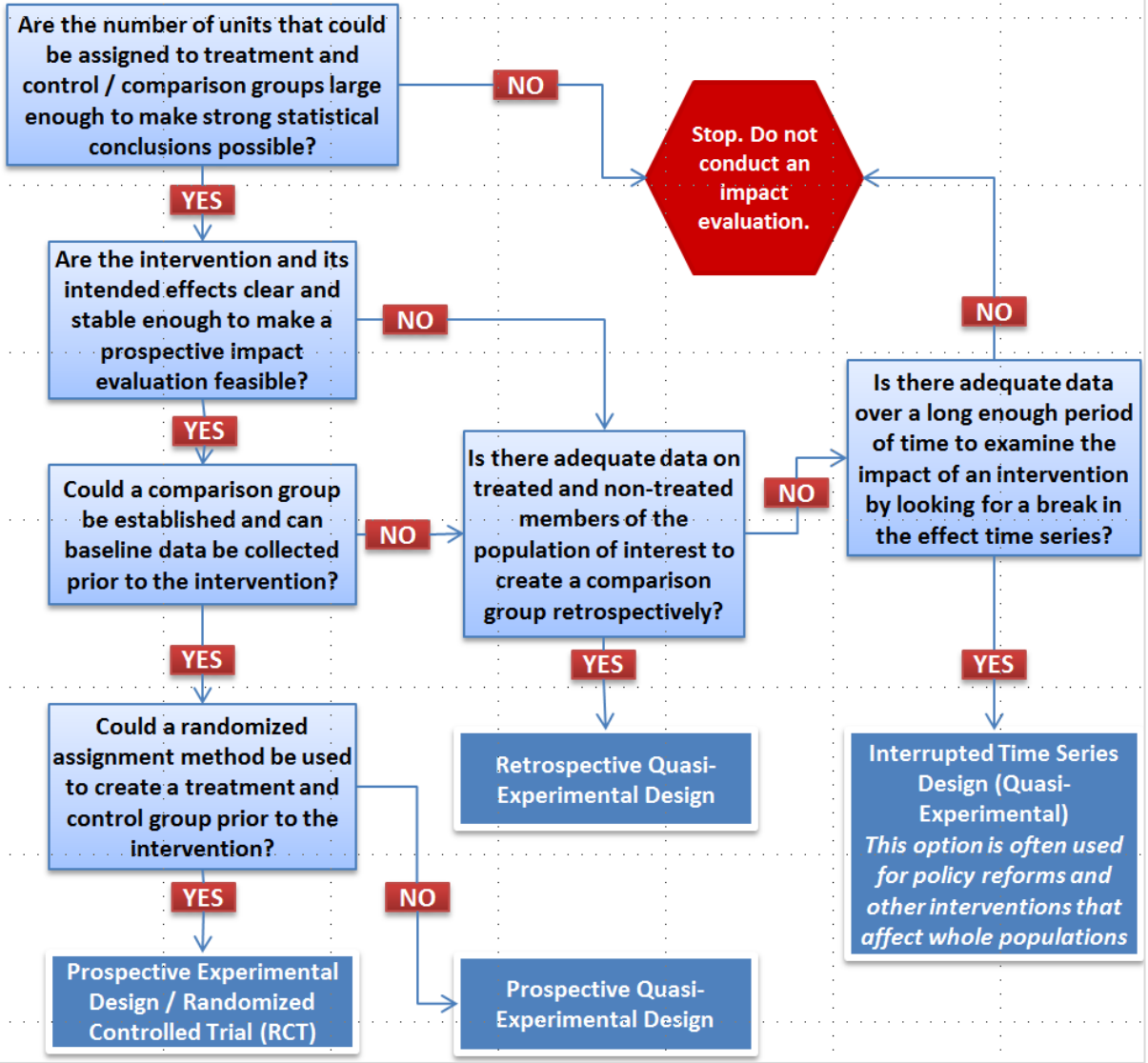
Strategic Considerations

- Will an impact evaluation help influence decisions about project design and / or resource allocation?
- Will impact evaluation results be available in time to inform decisions?

OR

- Is there a strong accountability demand for the evaluation?
- Is there buy-in from stakeholders such as implementing or scale up partners?
- Is there enough expertise, either in-house or through external help, to design and manage the evaluation?

Design Decision Chart*



Adapted from the Project Starter toolkit developed by Carana Corporation for the Office of Trade and Regulatory Reform (E3/TRR)