USAID
FROM THE AMERICAN PEOPLE

GUIDE FOR PLANNING LONG-TERM IMPACT EVALUATIONS (LTIEs)

Utilizing the Expertise of the Expanding the Reach of Impact Evaluation (ERIE) Program Consortium

September 2018

This publication was produced at the request of the United States Agency for International Development. It was prepared independently by the ERIE consortium.

TABLE OF CONTENTS

LIST OF ACRONYMS

ADS	Automated Directives System
ASTER	Advanced Spaceborne Thermal Emission and Reflection Radiometer
AVHRR	Advanced Very-High Resolution Radiometer
BRIGHT	Burkinabe Response to Improve Girls' Chances to Succeed
CARE	Cooperative for Assistance and Relief Everywhere
CEGA	Center for Effective Global Action
CHOICE	Child Health Opportunities Integrated with Community Empowerment
CIFF	Children's Investment Fund Foundation
DHS	Demographic and Health Surveys
DID	Difference-in-Difference
EOS	Earth Observing Systems
ERIE	Expanding the Reach of Impact Evaluation
GFC	Global Forest Cover
GIS	Geographic Information System
IRB	Institutional Review Board
ITS	Interrupted Time Series
IV	Instrumental Variables
LIDAR	Light Detection and Ranging
LTIE	Long-Term Impact Evaluation
M&RI	Measles and Rubella Initiative
MCC	Millennium Challenge Corporation
MERLIN	Monitoring, Evaluation, Research and Learning Innovations Program
MODIS	Moderate Resolution Imaging Spectroradiometer
NDA	Non-Disclosure Agreement
NDIGD	Notre Dame Initiative for Global Development
PALSAR	Phased Array L-Band Synthetic Aperture Radar
PCI	Project Concern International
PPTAL	Brazil Indigenous Lands Project
PSS	Post-Project Sustainability Study
PSUR	Programa de Sostenibilidad y Unión Regional Sur
RCT	Randomized Controlled Trial
RD	Regression Discontinuity
SAR	Synthetic Aperture Radar
SIA	Supplemental Immunization Activity

EXECUTIVE SUMMARY

International development programs are designed to achieve important improvements in health, education, and financial well-being for disadvantaged populations (e.g., reduced maternal and child mortality; increased farmer crop yields; or improved primary education enrollment rates). Despite these intentions, many development programs do not measure whether they ultimately achieve their desired results because the outcomes can be challenging to measure, and may occur well into the future. As a result, we often miss out on opportunities to learn about the most effective and efficient uses of scarce development resources.

An impact evaluation measures program effectiveness by comparing the outcomes of those who received a program against those who did not. By establishing a valid counterfactual, or an estimate of what an outcome would be in absence of the program, impact evaluations can credibly measure and attribute any changes in outcomes between participants and non-participants to the program. With this information, decision-makers can make more informed choices about which programs to support and how to improve them.

A long-term impact evaluation (LTIE) measures outcomes over longer time horizons than a typical impact evaluation. LTIEs typically occur after program implementation is complete, and address questions about a program's impact that cannot be answered within the span of a typical evaluation or program life cycle. They can provide crucial information on whether observed outcomes are sustained over time, allowing the United States Agency for International Development (USAID) and others to make long-term strategic management and funding decisions, particularly in sectors with theories of change with critical outcomes materializing well after a program ends.

To develop opportunities to measure long-term impact, USAID launched the **Expanding the Reach of Impact Evaluation (ERIE)** initiative. This resource guide will help managers identify programs that are candidates for LTIEs, understand potential LTIE designs, and prepare to conduct them. The guide also provides information on traditional impact evaluation designs and data collection methods; explains how to apply them to long-term evaluations; gives guidance on the use of traditional and innovative data sources; and provides several examples across a range of sectors. Program managers can make decisions about measuring long-term impacts in three main scenarios:

1. **Setting up an LTIE for a new program.** Programs may decide to put in place LTIE designs from the onset of a program if decision-makers anticipate that outcomes will be realized over a long-term time horizon.
2. **Following up on a previously conducted impact evaluation.** LTIEs can be implemented as follow-ups to previously conducted impact evaluations of the same program.

3. **Developing an LTIE design ex-post.** Program managers may be interested in generating evidence about program effectiveness from activities that did not previously measure impact, in order to make decisions about continuing, replicating, or scaling the program. Though it is generally more challenging to implement this form of LTIE compared to the previous two types, it is sometimes possible to retroactively identify or reconstruct a comparison group that did not receive the intervention.

There are certain prerequisites to making an LTIE possible. The first is having a theory of change that projects long-term impacts and the underlying logic of how they will be realized. A primary consideration for investing in LTIE is whether the program's theory of change, or project logic, anticipates that it will achieve important outcomes after the program has ended. Programs that expect its most critical outcomes to occur after implementation is completed could be good candidates for LTIE.

The next prerequisite is being able to identify a credible comparison group and a way to measure the treatment and comparison groups over time. Evaluators can work with program implementers to examine whether an appropriate comparison group exists, and determine a data collection plan to track treatment and comparison groups for the full duration of the evaluation. The data collection plan should document existing data sources the evaluation will use as well as sources the evaluation will need to generate on its own. This guide covers the typical data sources used in LTIEs, which include both traditional data sources and new data technologies that can expand the options for these evaluations:

1. **Traditional data sources:** household surveys, focus group discussions, key informant interviews, national surveys, administrative data, and monitoring data.

2. **Innovative technologies:** satellite and other geospatial data and mobile data, which, in particular circumstances, allows for the analysis of exceptionally large quantities of data.

Finally, the last section of the guide provides **practical examples through overviews of LTIEs** conducted in areas including education, early childhood development, health, natural resource management, and land rights.

For more information on ERIE and whether to conduct an LTIE for your program, contact:

Email: MERLIN@USAID.gov
Phone: (202) 712-0000
Fax: (202) 216-3524

USAID.gov/GlobalDevLab/MERLIN/ERIE

I. INTRODUCTION TO THE GUIDE

Achieving critical change in international development outcomes often takes sustained effort over long periods of time. Indeed, many international development programs are designed to achieve long-term impacts, but evidence of these impacts is often unavailable because we rarely measure the effects of these programs over a long time horizon. Evaluations tend to focus on short- to medium-term outcomes for a number of reasons. Rigid program cycles often require that funding for evaluations align with program implementation, while insufficient evaluation budgets result in unreliable funding for follow-up evaluations after program completion. Without evidence of long-term impacts, development organizations do not have reliable information to make cost-effective decisions, particularly in sectors with theories of change that predict that outcomes will materialize well after program completion.

To develop opportunities and approaches for conducting more evaluations of a program's long-term impact, USAID launched the ERIE initiative. ERIE is one of five concepts within USAID's **Monitoring, Evaluation, Research and Learning Innovations Program (MERLIN)** program, which promotes the co-creation and co-design of measurement approaches providing innovations for monitoring, evaluation and learning. ERIE aims to conduct several pilots of long-term impact evaluations for completed USAID programs using innovative data collection strategies and methods to conduct these evaluations. ERIE also helps programs that are just starting up to put in place strategies to measure long-term impacts. These evaluations will examine whether any short-term impacts have been sustained, or measure other anticipated impacts that materialize over a long-term horizon. They will also help to develop the approach and generate lessons for other programs interested in measuring long-term impact.

The purpose of this guide is to help managers identify programs that are potential candidates for long-term impact evaluations, work with evaluators to design and conduct the evaluations, and prepare new programs or strategies for evaluations of long-term impacts. The guide provides information on traditional impact evaluation methods and data collection approaches, explains how to apply them to long-term impact evaluations, gives guidance on the use of traditional and innovative data sources, and describes several examples of long-term impact evaluations in practice. Certain portions of this guide may prove more useful to managers, who are looking for answers to the following questions:

- **What is long-term impact evaluation?**
- **Is long-term impact evaluation right for my program or activity?**
- **What should I do to prepare my program and/or activity to be evaluated in the long-term?**

What is Long-Term Impact Evaluation?

While the designation of long-term impact evaluation can vary depending on the sector or the program, LTIEs typically take place after a program has completed implementation, and they attempt to measure the ultimate outcomes of a program's theory of change. LTIEs are retrospective in nature, addressing questions about a program's impact that cannot be answered within the span of a typical evaluation, including whether short- or medium-term impacts observed in the program cycle are sustained over time, or whether long-term outcomes materialize as expected according to the program's theory of change. A common example of programs with long-term impact is effective early childhood interventions, which have important impacts on a child's life into adulthood. Measuring these long-term impacts has provided a solid case for governments to continue and even expand their programs (Garces, Thomas, & Currie, 2002). Box 1 below contains other examples of international, long-term impact evaluations:

Box 1: Examples of International LTIEs

Health

- AidData researchers evaluated a large-scale measles vaccination effort carried out in 26 sub-Saharan African countries using Demographic and Health Survey (DHS) data to identify impacts on child mortality, fertility, and educational outcomes. The evaluation, using a child-level regression model and a duration model, measured impacts up to eight years after the initial vaccination campaigns ended, and found that the vaccination substantially reduced child mortality and that subsequent maternal fertility also declined (BenYishay & Kranker, 2015). For more details on this evaluation, see page 44.
- Notre Dame Initiative for Global Development (NDIGD) researchers evaluated a USAID program designed to improve access to primary health care in Indonesia. The Child Health Opportunities Integrated with Community Empowerment (CHOICE) project trained caretakers and educated local community members on basic health care. Using mobile data collection to survey beneficiaries and a comparison group seven years after the program ended, the evaluation found that CHOICE led to increases in health care utilization for infants and young children, which was attributed to the number of skilled health-workers attending births. For more details on this evaluation, see page 46.

Education

- Mathematica researchers evaluated a girls' education project in Burkina Faso that constructed rural schools and supported complementary interventions

to improve girls' academic outcomes. The evaluation, which used a regression discontinuity approach, followed up on treatment and comparison groups at several intervals. Three years after the project started, the study found that the program increased school enrollment and test scores, while the seven- and ten-year follow-ups found that impacts on enrollment and test scores were sustained, especially for girls. The seven-year follow-up also found positive impacts on child labor, and the ten-year follow-up found increased primary school completion rates and reduced early marriage rates for girls (Davis et al., 2016). For more details on this evaluation, see page 37.

Land Tenure and Deforestation

- AidData researchers collaborated with the German Development Bank to conduct a large-scale evaluation of land rights and enforcement on deforestation rates in indigenous communities in the Brazilian Amazon eight years after program completion. Using a combination of data on community land boundaries; administrative records on the program criteria and timing; and high-resolution satellite imagery, the evaluation found no effect on deforestation, even for communities that received support for surveillance and enforcement (BenYishay et al., 2016). For more details on this evaluation, see page 48.
- Geospatial Impact Evaluation (GIE) researchers worked with the USAID Ecuador mission to evaluate the impacts of land titling programs on deforestation in the 5-10 years after the program ended. Using geospatial data on land cover, population density, proximity to infrastructure, and other variables that predict deforestation, the research team identified comparable areas that did not receive the land titling program. The study found that the land titling program had not significantly affected deforestation during the time period considered in the study, pointing to the need to link land titling interventions with other programs that promote forest conservation (Buntaine, Hamilton, & Millones, 2015). For more details on this evaluation, see page 41.

Climate Change and Migration

- Center for Effective Global Action (CEGA) researchers analyzed environment-induced migration in Indonesia and extrapolated lessons about migratory patterns. Using household panel data from a 15-year span (1993-2007), the researchers compared migration patterns to data on precipitation, temperature, and natural disasters to estimate the influence of environmental factors on migration (Bohra-Mishra, Oppenheimer, & Hsiang, 2014). By using long-term

data, the research team was able to analyze a number disaster events and weather cycles in all parts of the country within a long time frame, allowing for a much larger sample size, and to observe whether migration was temporary or permanent. They found that climatic variations influenced the rate of permanent migration, with an increasing rate of migration associated with temperatures over 25° C, while natural disasters had less or no impact on permanent migration.

Agriculture

- CEGA researchers used satellite data to estimate smallholder farmer crop yields in Kenya and Uganda. They developed new statistical techniques to generate yield predictions, which were then validated using field level data, including farmers' plot boundaries, and yield surveys. The technique is now being extended to India and Rwanda. As remote sensing techniques continue to evolve, they will allow long-term evaluations to incorporate up-to-date data, minimizing the need for survey infrastructure. For example, if a past agricultural intervention kept track of GPS coordinates of a farm, a comparison of current day agricultural yields might help estimate long-term effects of the intervention.

Two recent developments make long-term evaluations more feasible now than in the past. First, the recent increase of traditional impact evaluations has created a pool of well-documented, rigorously studied programs to which we can return for long-term follow-up. Second, a rapid improvement in the granularity, coverage, and nature of data has created new opportunities for rigorous, quasi-experimental evaluations. For example, advances in mobile data collection and the availability of "big data" sources allow evaluators to construct measures of key outcomes, such as poverty reduction or economic activity. Additionally, geocoding, data extraction tools, and satellite imagery allow us to define what activities took place (as well as where, when, and by whom). When the data sources are merged, they can create new opportunities for evaluation of development interventions that currently lack evidence of long-term impacts.

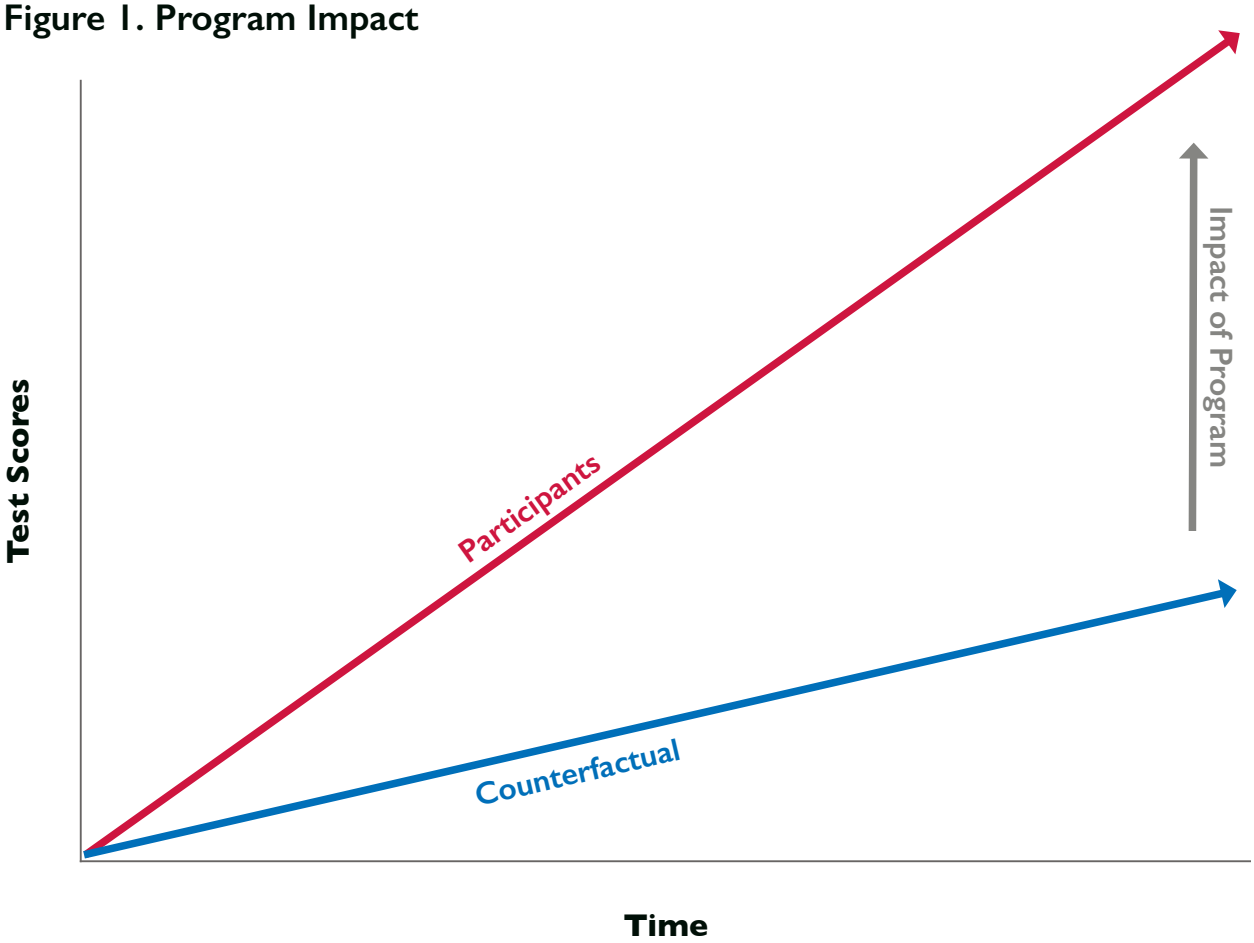
The remaining sections in this guide will introduce impact evaluation methods and their application(s) in order to prepare you for the section that covers LTIEs and typical data sources. The guide has been structured for ease of use by USAID missions and program managers. Section 2 defines impact measurement and long-term impact evaluation, in addition to providing an overview of the methods used to conduct these evaluations. Section 3 covers the data sources that can be used in long-term impact evaluations. Section 4 describes three basic types of long-term impact evaluations: 1) Evaluations that start with the program, 2) Following up on an existing design, and 3) Conducting evaluations after the program has ended. Section 5 covers how to prepare for disseminating and applying the results of the long-term impact evaluation, and Section 8 provides 1-2 page descriptions of LTIEs across a variety of sectors.

II. IS LONG-TERM IMPACT EVALUATION RIGHT FOR MY PROGRAM?

What Do We Mean by Impact?

Before describing how to conduct long-term impact evaluations, we need to describe what we mean by impact and how it is typically measured. Impact evaluation is different from other types of program evaluations because it not only measures the changes in outcomes over time, but is specifically designed to determine to what extent those changes can be directly attributed to the program. Impact evaluations do this by comparing outcomes for people offered the program with estimates of what their outcomes *would have been* had they not received the program. The "outcomes that would have been" is also referred to as the counterfactual scenario (Rubin, 1974). To give an example, Figure 1 below graphs the impact of a hypothetical education program by displaying the outcomes (test scores) for potential beneficiaries (red arrow), and what the scores would have been absent the program (blue arrow). The difference between the two arrows (gray arrow) is the impact of the program. Note that in this example, the people who did not receive the program still experienced improved test scores over time, but not as much as the beneficiaries who were offered the program.

Figure 1. Program Impact



Directly measuring the counterfactual scenario is not possible. Instead, evaluators attempt to estimate what would have happened by identifying individuals or communities who are similar to those offered participation in the program. This group not offered the program is known as a **comparison group**[1], while the group of participants is called the **treatment group**.

USAID's Automated Directives System (ADS) 201 defines **impact evaluations** as evaluations "that measure the change in a development outcome that is attributable to 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." (USAID, 2013)[2]. The goal of impact evaluation is to measure the specific contribution of the intervention to the observed change in outcomes. Other differences should be eliminated either through the careful selection of equivalent comparison units or by statistical adjustment for remaining differences that cannot be eliminated through the selection of equivalent units[3].

The difference between a traditional impact evaluation and long-term impact evaluation is that LTIEs measure outcomes over a longer time horizon. USAID program cycles are often three to five years. When impact evaluations are incorporated into the program, they typically measure outcomes that materialize during program implementation, or immediately after. In contrast, LTIEs would measure the impact on the same or other outcomes well after implementation of the program has ended. This type of evaluation provides critical information on whether observed changes are sustained over time, or whether other anticipated, long-term, outcomes result from the program.

Understanding a program's long-term impact can have many benefits, but making the investment in an LTIE may not be right for all programs. Determining whether to pursue an LTIE involves several considerations that we highlight below:

- **Does the program's theory of change anticipate long-term impacts?** A primary consideration for investing in a long-term impact evaluation is whether the program's **theory of change** anticipates that it will achieve important outcomes after the program has ended. Programs that are likely to end implementation before they can measure important outcomes—or programs that target outcomes for which the time horizon of change is lengthy—may be good candidates for LTIE. Box 2 on the next page provides more information on program theory of change and its main components.

[1] In the case of randomized controlled trials, this group is the control group.

[2] For more information on when to use impact evaluation methods and their designs, please see USAID Technical Note: Impact evaluations. https://www.usaid.gov/sites/default/files/documents/1870/IE_Technical_Note_2013_0903_Final.pdf

[3] The definition of a "unit" varies depending on the type of intervention. In the education sector, a unit might be a teacher, a student, or a school. In agriculture, units could be defined as individual farmers, households, or farmer cooperative associations.

Box 2: Program Theory of Change (An Example)

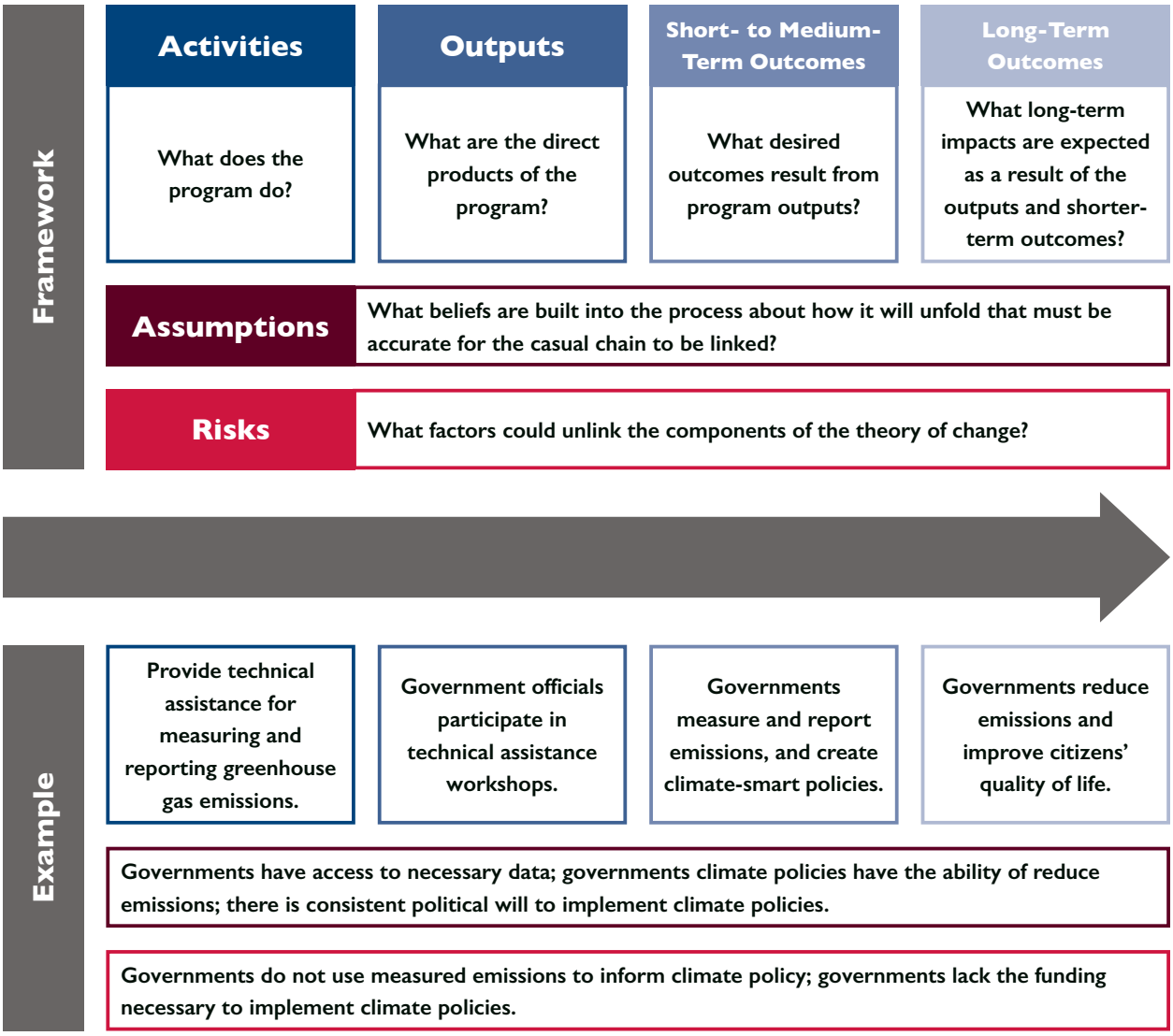
A theory of change for a specific program or intervention describes the anticipated causal pathway from a program's activities to its intended effects, including its direct outputs as well as its short and long-term outcomes. Theories of change also include the key assumptions about how the program produces its outcomes as well as indicators to measure progress (USAID, 2017). A logic model is a graphical representation of the theory of change that summarizes its key elements and illustrates the links between the main program inputs, activities, outputs, and, ultimately, the targeted outcomes[4]. Figure 2 on the next page provides an example logic model for a program, funded by the Children's Investment Fund Foundation (CIFF), that is designed to have long-term impacts (CIFF, n.d.).

With the long-term goal of reducing emissions and improving quality of life for residents in urban areas, the program delivers technical assistance and tools to governments (the activities) through workshops (the output) to help them better measure their greenhouse gas emissions (the short-term outcome), and develop climate-smart policies (the medium-term outcome). The program assumes that governments will have access to the data required to measure emissions, the political will to set better climate policies, and that those policies have the ability to reduce emissions. In addition, risks of the program's failure are that cities may not have the financial resources to enact effective emissions-reducing policies (CIFF, n.d.).

- **Is measurement of a program's cost-benefit important for decision-making?** Cost benefit analysis is a systematic process to determine the overall value of a program or intervention to society by comparing its benefits to the cost of delivering and maintaining it. In short, a program whose benefits outweigh its costs can be considered a good investment especially if its net benefits are larger than other programs with similar goals. Measuring the long-term impact of programs through LTIE can greatly improve the accuracy and reliability of cost-benefit analyses.
- **Will results from the LTIE influence decision-making?** Understanding a program's long-term impact and how benefits manifest over time can help implementers think not just about whether to expand a program, but also about the required coverage that a program needs to achieve and sustain impacts. For example, deworming programs in countries where soil-transmitted helminths and schistosomiasis are prevalent are highly effective when delivered through schools (Miguel & Kremer, 2004). In addition, a long-term follow-up conducted 10 years after the end of a deworming experiment found

[4] For more information on how to develop a theory of change and logic model, please see this How To Note: Developing a Project Logic Model: https://usaidlearninglab.org/sites/default/files/resource/files/project_logic_model_how_to_note_final_sep1.pdf

Figure 2. Example of a Logic Model



that the program had positive impacts on employment and hours worked per week (Baird, Hicks, Kremer, & Miguel, 2016). These findings have led to increased funding for and the scaling of deworming campaigns around the world (Croke, Hicks, Hsu, Kremer, & Miguel, 2017).

- **Is an LTIE technically feasible?** For a long-term impact evaluation to be successful, it must be able to measure the counterfactual in a credible way. Each LTIE will need to identify a credible comparison group and track the key outcomes for that group over time. In the next section, we provide an overview of the methods that can be used to identify and measure treatment and comparison groups, including some of the opportunities and challenges in applying these methods for long-term impact evaluations.

III. SETTING UP A LONG TERM IMPACT EVALUATION

Program managers can make decisions about measuring long-term impact in a number of different scenarios. They may be at the starting point of a program and would like to put in place an evaluation plan that allows for long-term follow up after the program ends. Alternatively, they may have results from a previous evaluation, and are wondering if the impacts were sustained, increased or decreased over time. Finally, program managers may be interested in generating evidence about program effectiveness from activities that did not previously measure impact in order to make decisions about continuing, replicating or scaling the program. We describe these scenarios in more detail throughout the following pages in this section and give an overview of the conditions that need to be in place to make LTIE possible.

Scenarios for Long Term Impact Evaluations

Scenario 1: Setting Up an LTIE for a New Program, or an Old Program in a New Location

Programs may decide to put in place long-term impact evaluation designs from the start of a program when decision-makers anticipate a long-term horizon for achieving the main program outcomes. In this scenario, implementers integrate the long-term impact evaluation into the startup and implementation of the program. An important benefit of planning a long-term impact evaluation at the outset is that evaluators have a broader range of design options, including experimental designs. To plan an LTIE at the beginning of implementation:

- Program decision-makers must first identify the specific long-term outcomes the program aims to achieve, the time period needed to achieve the outcomes, and how these outcomes can be measured. The process of developing a program theory of change (described previously in Section 2) should also help to identify important program milestones and what to measure at all stages of implementation as the program delivers on its targeted outcomes.
- After making the decision to develop a long-term impact evaluation design, program decision-makers should collaborate with evaluators to develop the specific research questions and explore the technical feasibility of the design. In particular, evaluators will work with implementers to identify and track an appropriate comparison group that will serve as the counterfactual in the impact analysis. The team would also have to consider how evaluators could collect data on the treatment and comparison groups after the program has ended and over the long-term.
- Evaluators then work closely with implementers to develop a plan for collecting data on outcomes for treatment and comparison groups in order to collect data at the appropriate time points according to program's theory of change. The data collection

Box 3: Key Considerations for Planning a Long-Term Impact Evaluation for a New Program

Counterfactual: Determine whether there a reasonable likelihood that the evaluation can track a both the participant and comparison groups over the full course of the evaluation, even after the program has ended.

Data Collection: Develop a strategy for tracking treatment and comparison groups over the long-term. Document program implementation so that it can be used in the final evaluation. Ensure consent for survey data collection (if relevant) allows for long-term follow-up.

Evaluation Management: Identify who will manage the evaluation after implementation of the program ends.

plan should anticipate which data sources will be available and which sources the evaluation will need to generate on its own.

- If the evaluation requires surveys, the evaluation design process will include assessing whether the research team can survey the same sample of individuals, households, or other units over time, or whether they will collect data from distinct but equally representative samples within the treatment and comparison groups at the different points in time. Evaluators will also need to track changes in the program design, participants, communities, and other contextual factors that may influence an evaluation or its findings. These contextual factors will require a longer period of program monitoring than a typical evaluation.

Challenges

One of the main challenges in setting up a long-term impact evaluation before the program starts is determining who will manage the evaluation over the long-term and how it will be funded. Some donors may have evaluation funding that is separate from program funding while others do not. In most cases, a long-term impact evaluation may fall outside the program's funding period, and therefore may be difficult to plan for and allocate resources to in advance. In these instances, the program may fund an impact evaluation design that will deliver interim results before the end of the program cycle with the hope that future programs or evaluation units will take on a long-term follow up. In other cases, the program may be able to develop external partnerships with foundations or organizations who can fund and implement the evaluation after the end of the program.

Scenario 2: Following Up On a Previously Conducted Impact Evaluation

LTIEs can follow up on previously-conducted impact evaluations of the same program to measure its long-term outcomes. In many cases, the study will collect new data on the treatment and comparison groups and then assess whether program impacts have been sustained over time, whether impacts increased or decreased, or whether new impacts materialized. Follow-up LTIEs involves:

- A careful review of the previous impact study to ensure that there is sufficient information about the design and data collection so that evaluators can conduct a follow-up evaluation. Typical materials for this review include the evaluation final report and data used in the analysis as well as clear documentation of the data sources, implementation records, and the pre-analysis plan (if available). The pre-analysis plan should indicate the specific outcomes the evaluation was designed to measure[5].
- A thorough examination of the data from the previous impact study to ensure there is sufficient data quality and that appropriate data management techniques were followed to enable longitudinal and comparative statistical analysis.
- Working together with program decision-makers, the research team selected to conduct the follow-up will refine the specific research questions and the long-term outcomes the study will measure. This would likely include a close review of the program's theory of change, assumptions, and a review of program implementation.
- Gathering data on the treatment and comparison groups. This could involve conducting primary data collection or using other secondary or satellite data sources. Finally,

Box 4: Key Considerations for Following Up On an Existing Evaluation Design

Counterfactual: Review information from the previous study and assess whether the comparison group in the original design remains valid.

Data Collection: Assess program implementation for consistent delivery and reach. Develop a plan to locate the originally selected treatment and comparison groups, and ensure that the follow-up data collection is approved by an institutional review board (IRB).

Evaluation Management: Determine whether to work with an earlier evaluation team on the follow-up evaluation.

[5] This information may also be included in the evaluation design report, or in other program documentation.

researchers would analyze the data, write up and help to disseminate the results.

Challenges

The greatest challenge to following up on an existing design is ensuring that comparison group would continue to represent a valid counterfactual for the treatment group. For example, if the comparison group were offered the program shortly after the original evaluation ended, the follow-up evaluation would be comparing different lengths of program exposure rather measuring the impact of participating in the program. In that case, the evaluation team would need to determine if there is another credible comparison group that can be identified. For example, if an education intervention was offered in specific grades, it might be possible to identify similar groups of students even if all of the students in a particular grade eventually received the program.

Scenario 3: Developing an LTIE Design Ex-Post

Many types of interventions are designed and implemented without a comparison group in mind. This does not make ex-post long-term evaluation impossible, but it does raise a number of important challenges. The most critical of these is identifying or reconstructing a group of comparison units that did not receive the program. It is possible to do this with a variety of techniques, but all depend on having information about the criteria and process used to select participating units into the program. The evaluation will pair this information with relevant background data of non-participating units to identify a comparison group. Confidence about the evaluation results will largely depend on the credibility of the comparison group and its equivalence (in observable and unobservable ways) to the units selected for the intervention. The most common steps for conducting an **ex-post** LTIE are:

- A careful review of the program's theory of change to understand both the targeted outcomes and the time period over which the program would have an impact on those outcomes. In some cases, the program may not have documented its theory of change or results framework, and the evaluation will need to reconstruct the theory of change ex-post to determine the key research questions and most appropriate design. Key questions to ask are:

    1. What immediate steps were realized during the project itself? Were outputs achieved as expected? Were any outcomes or short-term results realized or measured?

    2. What long-term impact could reasonably be detected at this point, taking into account the type of impact and the time that has passed since project end?

    3. What process and criteria were used to select participants? Were there criteria that determined priority in terms of being admitted into the program?

Box 5: Key Considerations for Designing Ex-Post Long-Term Impact Evaluations

Counterfactual: Is it possible to identify a valid comparison group using program selection criteria? Were these selection criteria consistently applied?

Data Collection: Are there data sources available that can serve to establish the baseline and endline for treatment and comparison groups? Is there sufficient information on program implementation or the possibility of working with implementers of the previous program?

Evaluation Management: Ensure that the evaluation team has significant experience with the data sources to be used. Ex-post LTIE designs can be commissioned by program decision-makers or by new users of the evaluation results.

- Determine the data the evaluation will use to analyze impacts, in particular those outcomes that might not be observable during the time frame of more traditional impact evaluations. Review existing data sources for the program, as well as the participant and potential comparison groups, the timeframe for these data, and their periodicity.
- The final evaluation design will depend primarily on the specific research questions the evaluation seeks to address as well as the amount of information available about program selection, coverage, potential comparison groups, and ability to measure the key outcomes. As described in the previous sections, there may be many innovative ways to collect data for treatment and comparison groups, including remotely sensed measures and long-term administrative data[6].

Challenges

The main challenge to ex-post LTIE designs is having confidence that the only difference remaining between the treated group and the comparison group is the presence of the intervention. Because ex-post evaluation designs are able to select the comparison group using observable characteristics only, it is possible that there are unobservable differences between the treatment and the comparison groups that are also correlated with the outcomes the evaluation is trying to measure. Evaluators often conduct sensitivity analyses to examine how impact estimates might change in the presence of unobserved characteristics[7].

[6] Administrative data refers to data that are collected for purposes other than research. More details on the definition can be found on page 26.

[7] Sensitivity analysis is a method used to learn how uncertainty in a mathematical model can affect the outcome of interest. For more on sensitivity analysis and the use of observable control variables, see Oster 2017.

IV. METHODS FOR LONG TERM IMPACT EVALUATION

As mentioned previously, the same methods used to identify a credible comparison group in traditional impact evaluations are also used in long-term evaluations. Throughout this section on the following pages, we offer a brief overview of these methods and how they would apply to LTIEs. While the specific statistical techniques for these methods are beyond the scope of this guide, there are many resources that cover this information[8].

Experimental Impact Evaluations

Experimental impact evaluations are conducted through **randomized controlled trials** (RCTs). They identify a credible counterfactual by randomly assigning eligible units to treatment and comparison groups prior to their participation in the program, and then track both groups over time. These eligible units could be defined as individuals, villages, schools, or other units. With a large enough sample, well-conducted RCTs with truly randomized assignment ensure that the comparison group is the same as the treatment group, in both observable and unobservable ways, with the exception that the treatment group is offered the program. All RCT designs require randomly assigning eligible units prior to starting the program, while LTIEs using this design would need to check in periodically on the treatment and comparison groups over a long period of time to ensure they remain distinct. Because it may be difficult to track specific individuals, some LTIEs track outcomes at the community or other geographic or organizational level.

Box 6: Example of a Long-Term Experimental Impact Evaluation

The U.S. Department of Education's Upward Bound program is one of the largest and longest-running federal programs designed to help disadvantaged students prepare for, and attend college. In this study, evaluators randomly assigned eligible students to treatment and comparison groups, and followed up with students seven to nine years after their expected high school graduation date. While the study found positive impacts on high school math credits and the likelihood of earning a postsecondary certificate or license from a vocational school, there were no detectable effects on other outcomes, including graduation and grades; postsecondary enrollment; financial aid application or receipt; or the completion of a college degree (Seftor, Mamun, & Schirm, 2009). However the U.S. Department of Education continues to fund Upward Bound (United States Department of Education, n.d.).

[8] The World Bank's Impact Evaluation In Practice is one very good example: http://www.worldbank.org/en/programs/sief-trust-fund/publication/impact-evaluation-in-practice

Quasi-Experimental Impact Evaluations

Quasi-experimental impact evaluations differ from experimental evaluations in that they identify a comparison group **not through randomization** but by statistically matching comparison units with the treatment group or by tracking parallel trends between the treatment and comparison units. These types of approaches are often appropriate when randomization is not feasible due to practical or ethical reasons. There are a number of methods to ensure rigorous matching in this type of evaluation:

- **Matched comparison group designs** offer several methods for identifying a potential comparison group, often using the eligibility criteria for program participation. These may include location, demographic characteristics, and pre-program values of outcomes of interest. Matching is common for long-term impact evaluations conducted after the program has started (or even after the program has ended). A successful LTIE using a matching design requires having data on both participating units and potential comparison units, on pre-intervention outcomes and important variables that are likely to influence participation in the intervention. As with all matching methods, identifying a credible comparison depends on the characteristics that are used in conducting the match. One type of a matched comparison group design is **propensity score matching**, which creates a comparison group by estimating the probability that participating and non-participating units would enroll in a program based on observable characteristics relevant to enrollment. For example, NDIGD researchers conducted a seven-year impact evaluation of the Millennium Challenge Corporation (MCC) water intervention in Ghana using PSM methods. The variables used to find a comparison community which matched the communities where improved water facilities were installed included: **water-system related variables** (presence, quality, distance of current water systems, prevalence of waterborne diseases), **physiographic variables** (elevation, rainfall, land cover, agricultural growing periods), and **demographic variables** (poverty index, population size) (Guzman, Brown, & Khatiwada, 2016). The propensity score is computed for all units and used to estimate the counterfactual by matching comparison to treatment units (Gertler, Martinez, Premand, Rawlings, & Vermeersch, 2011).

- **Synthetic control** can be used to estimate the impacts of macro-level or large scale policy impacts. The method involves creating a counterfactual that most resembles the treated unit prior to implementing an intervention or policy by using a weighted average of several units that share similar characteristics to the intervention group (Abadie & Gardeazabal, 2003). For example, to evaluate the impact of a private sector reform in Rwanda, researchers created a "synthetic Rwanda" by using several pre-reform macroeconomic characteristics from seven other countries that did not implement the reform, such as trade balance and GDP per capita, to create a weighted average that was then compared the outcome of interest in Rwanda . The authors found a 186% average increase in new firms approximately one year after the reform was introduced (Gathani,

Stoelinga, & Santini, 2013).

- **Quasi-experimental difference-in-difference (DID)** identifies a similar population to that of the treatment group, and estimates the impact of a program by comparing the change in outcomes for the treatment group to the change in outcomes for the comparison group over time. With data on key outcomes for both groups before and after the program has started, evaluators estimate program impact by calculating the difference in key outcomes between the baseline and endline, and then comparing the two differences for the treatment and comparison groups. DID designs can be useful when baseline outcome levels differ between the treatment and comparison group. However DID requires having baseline outcome values for the treatment and comparison groups, and must satisfy the assumption that outcomes for the comparison and treatment groups followed the same trajectory before the program started. The DID approach can use a wide variety of data sources beyond survey data, such as satellite imagery and administrative records. Data collection methods for LTIEs are discussed in greater depth in Section 5.

Instrumental Variables

An instrumental variables (IV) approach uses an external source of variation, called an instrumental variable, to determine participation in an intervention (Gertler, Martinez, Premand, Rawlings, & Vermeersch, 2011). Evaluators can use this method to measure impact when they can identify an IV that influences who participates in a program, but it is outside of the control of eligible participants and influences participant outcomes only through the program. Examples of instrumental variables that have been used in development program evaluations include distance to different types of infrastructure (for an example, see Box 9). IV can be a useful tool for LTIE when geographic features that affect exposure to a particular intervention allow evaluators to identify causal long-term impacts. For example, researchers evaluated the effects of hydropower electrification on local economic development in Brazil from 1960 to 2000 using variation in

Box 9: Example of a Long-Term Evaluation Using IV

To estimate the impact of a newspaper-based campaign in Uganda that informed communities of funding allocated to their districts for schools, researchers compared schools with greater exposure to the campaign to schools with less exposure to the campaign. Using distance from a school to the nearest newspaper outlet as an instrument for the level of teacher exposure to an information campaign, the study found that six years after the campaign began, schools with greater exposure to the campaign received a larger share of the total funds allocated to the district (Reinikka & Svensson, 2010).

electrical grid expansion as the IV (Lipscomb, Mobarak, & Tania Barham, 2013). They found large positive effects of electrification on a measure of human development and on housing prices (Lipscomb, Mobarak, & Tania Barham, 2013).

Regression Discontinuity

Regression Discontinuity (RD) is a evaluation method that uses a program's clearly defined selection criteria to estimate the impact of participating in the program. For this method, the treatment and comparison groups include the people or other units which fall just above and just below the cutoff point. (Common examples eligibility criteria include wealth indices or test scores.) In contrast to other impact evaluation methods, RD measures impact only for those who are near the cutoff point, rather than the average impact for all eligible participants. Unlike RCTs, RD does not require that an impact evaluation be planned in advance of program implementation, but the design requires having sufficient data to establish the trend in outcomes and establish the discontinuity. An LTIE using RD also requires that the strict, identifiable cutoff for participation eligibility remain active for a long enough period of time so that outcomes for the treatment group can be effectively compared to a comparison group that did not receive the program.

The impact evaluation of the Burkinabe Response to Improve Girls' Chances to Succeed (BRIGHT) Program (see page 37) provides an example of an LTIE that uses regression discontinuity design.

Interrupted Time Series

Interrupted Time Series (ITS) is a method that examines outcomes before and after an intervention clearly interrupts a trend in the outcomes of interest. ITS uses multiple data points over time before an intervention is introduced in order to establish the trend line. The trend line can then be used to estimate the counterfactual after a program is introduced. The "before" trend line is compared to the trend in outcomes for participants after the program has been implemented. The validity of this estimate depends on how well a pre-intervention trend can be established as well as the likelihood that the trend would continue. Evaluators using the approach will investigate whether the pre-intervention trend line is a credible counterfactual based on whether other external factors introduced or manifested within the same time period as the intervention could bias the impact estimates. Interrupted time series analysis requires having access to outcomes data from a large number of points in time to construct a time trend line that can serve as the counterfactual. In a long-term impact evaluation, it may be challenging enough to acquire any data before the intervention began and ensuring multiple rounds of data produces an even greater challenge.

A Note on Contamination, a Common Challenge for Long-Term Impact Evaluations

One challenge in conducting all long-term ex-post evaluations is to identify whether there has been **contamination *or* spillover** between the treatment and comparison groups. When comparison units end up participating in or being otherwise influenced by the program, then they can no longer serve as a valid counterfactual for the treatment group. For example, if the construction of regional market infrastructure has benefits for farmers outside the immediate target area, comparing the two sets of farmers would not generate an accurate estimate of program impact. It will be important, especially in a long-term impact evaluation where a longer period of time has elapsed since the program was implemented, to identify whether contamination of the comparison group has occurred. If the participants (or units) in the comparison groups have received, participated in, or otherwise benefitted from some aspect of the intervention then the comparison group cannot serve as pure control in measuring the impact of the program. In some cases, the impact evaluation design may be able to measure the contamination of spillover of a program by specifically tracking these non participants who likely benefited from the program, and comparing their outcomes to the outcomes for the larger comparison group.

V. DATA SOURCES FOR LONG-TERM IMPACT EVALUATION

LTIEs can use a wide variety of data sources for their analysis of results. New data technologies have expanded the opportunities for long-term evaluation, while more traditional data sources, such as survey or administrative data, continue to play an important role. Below we describe the typical data sources and give an overview of their use in LTIEs, and cover primary, secondary, geospatial, and mobile data sources. While geospatial data and mobile data could be considered secondary data sources in many instances, we note these separately because of their relevance for long-term impact evaluation.

Primary Data

Primary data for LTIEs is data that the evaluation will collect directly from the program or its beneficiaries, and can be qualitative or quantitative in nature. The collection of primary data involves hiring field teams to administer questionnaires, follow open-ended interview protocols with respondents, or other direct data collection methods. Sources for LTIEs include the following:[9]

- **Surveys** conducted as part of a long-term impact evaluation can gather **cross-sectional data**, which is data collected from a sample that is representative of the population at a specific point in time, or **longitudinal data**[10], which is data gathered from the same respondents or populations over time. Longitudinal household surveys, often used in impact evaluations, gather information on a long list of household characteristics such as health, attitudes, income, and consumption, and demographic information for each member. These surveys help understand and document changes in household wellbeing over time. Repeated cross sections can look at similar outcomes for the survey sample without tracking the same households. A long-term impact evaluation might conduct a follow-up survey several years after the program has ended to measure changes in outcomes, and would compare this data to either panel or cross-sectional data depending on the structure of the previous surveys.

- **Qualitative data** typically comes from interviews or meetings to gather **open-ended information** from respondents, and involves structured or semi-structured interview protocols. The open-ended format for interviews with individuals or focus groups allows respondents to answer questions in their own words, and allows the interviewers to adjust the questions based on information that they are receiving in the interview. Qualitative data might also include observations of individuals or groups in a setting of interest, such as a classroom or community meeting. Qualitative data collection can be

[9] There are other types of primary data such as data from observations or sensors for example, but these data sources are likely to be less relevant for LTIEs.

[10] Also called panel data.

used in a formative manner to refine the LTIE design, develop specific research questions, or develop survey instruments. For example, community perceptions or stories about a past program may indicate specific outcomes that the evaluation will measure. For example, perceptions of returns to girls' education could be explored through focus group discussions could help to develop specific measures for the evaluation.

Challenges to Primary Data Collection

Primary data collection can be costly compared to other data sources because it requires hiring field teams, developing instruments, and following data collection protocols to ensure the quality of the data. Primary data collection typically requires approval from an IRB, which is an impartial entity set up to ensure that the wellbeing of those who are participating in a research project is protected[11]. Research with human subjects also requires that researchers obtain informed consent from participants. This step involves informing research participants about how the data will be used and who will have access to it. LTIEs relying on primary data collection or data that was previously collected will need to ensure that the consent language used in previous data collection efforts allows for LTIE evaluation team use of the data and follow up of the respondents. New programs seeking to incorporate LTIEs can draft consent language that allows for long-term follow up from other research teams.

Primary data collection also requires that the field teams locate a sufficient number of targeted respondents to allow for credible analysis of their data. This may be especially challenging years after a program has ended because participants may have moved away, or are unreachable for other reasons. Programs setting up new LTIEs will need to determine how they will track the comparison and participant groups over time. If populations are not very mobile, the survey may be able to rely on GPS coordinates for the household. Alternatively, the study could collect and use cell phone numbers to follow up frequently with survey respondents. In limited cases, it may be possible to track individuals through administrative records, though privacy concerns make this a difficult strategy.

Secondary Data

Secondary data is a broad category of data that covers most types of data that have already been collected. This includes diverse types information such as general purpose data sets collected for a variety of analyses such as DHS or data that organizations collect in the process of running their programs. Because it does not require a dedicated data collection effort, secondary data is typically much less costly for the LTIE than primary data. However, evaluators working on impact evaluations need to check the quality of the data, its relevance to program outcomes, and its coverage in order to determine whether it will support credible impact measurement. Examples for LTIEs include:

[11] Research with human subjects is defined as a systematic investigation to develop or contribute general knowledge that involves obtaining information about living individuals through intervention or interaction with individuals is considered research with human subjects (United States Department of Health and Human Services, n.d.)

- **National surveys** that collect information from a nationally representative sample of households within a country on topics such as the population's health, education, and sociodemographic characteristics. National government agencies or other large institutions that conduct these surveys have often repeatedly collected the data over the course of many years, making them especially helpful for long-term impact evaluations by providing consistent data over long periods of time and across geographic locations. The main downside of nationally representative data is that because of their scale, they may not provide a sufficient number of data points to analyze data at the level of the target populations of individual programs. Analysis utilizing this type of data source may need to be conducted at a higher level, such as county or district.
- **Administrative data** are data that gathered, used, and stored for the purposes of implementing a specific program or policy. Examples include the eligibility criteria of program participants, individual take-up and use of a program, payment or tax information, specific products or services offered and received, program budget and spending, etc. They are especially useful to long term impact evaluation because they can provide a record of how the program was implemented and whether there were any variations in how it was delivered. In some cases, LTIEs may be able to use administrative eligibility criteria to identify credible comparison groups of individuals who would have been eligible for the program but did not participate.
- **Monitoring data** are data collected specifically to measure the performance of a program. For some programs, there may be an overlap between administrative and monitoring data when administrative data is used to assess program implementation. Examples of monitoring data include the number of program participants applying what they have learned in a training program, or participant satisfaction with the program. Like administrative data, monitoring data is especially important for LTIEs because it provides information on variations in program implementation and performance that helps evaluators interpret evaluation results. However, monitoring and administrative data are typically limited to the period of program implementation. They do not typically allow evaluators to study whether the program resulted in its desired outcomes or what happened to individuals who did not participate. These data sources can also inform the intensity of implementation for subgroups or individuals.

Challenges to Secondary Data

Secondary data are unlikely to contain all of the information needed to conduct an LTIE. In addition, the data quality of secondary data needs to be carefully assessed. Although national surveys and census usually publicize their data collection methodologies, administrative data sources do not. Monitoring data may not include information on data standards and can be of varying quality. Other challenges include data availability and meeting the timing needs of the LTIE. Because evaluators lack control over the type, timing, quality and availability of the data,

the data may not have sufficient coverage of the participation and comparison groups. As a result any secondary data source for an LTIE will likely need to be supplemented by other data sources. Finally, the use of secondary data sources may require special permission to use, and may not be fully available to the public.

Satellite Data

Satellite and other geospatial data offer new ways of conducting LTIEs, particularly of natural resources and infrastructure programs. Earth Observing Systems (EOS) generate remotely sensed data about almost anything that can be observed from space from satellites orbiting the earth and occasionally from aircraft and balloons. Satellite data can capture one-time occurrences, such as completed construction, whether a farming plot was cleared, or even if a crop transition occurred. In addition, satellite data can track continuous occurrence such as changes in greenness over a region or in nighttime light emissivity.

Evaluators can use remotely sensed data to measure impacts on a program's targeted outcomes as well as to identify control variables for the impact analysis. Examples of outcome variables from satellite data include crop or land-cover change, increases or decreases in population, improvements in water quality, or increased forest cover or decreased desertification. Examples of control variables include variables used in a 2014 impact evaluation of a land tenure program in Ecuador. Buntaine, Hamilton, and Millones (2015) used variables such as forest cover and loss, population, distance to major roads, electric grids, and other geographic features as control variables for the land tenure study. These variables are useful for identifying areas that are comparable to the areas where an intervention has taken place, both in terms of their environmental and social characteristics, as well as in terms of their temporal behaviour (seasonality, decreasing or increasing trend). In other words they are used to identify control groups or areas that serve as counterfactual to the "treatment" groups and areas.

Satellite data provides opportunities to measure certain program impacts that would be difficult to measure through traditional data collection. Its advantages include:

- **Information cost:** Several sources make satellite data available for free to certain users. Even when data need to be purchased from commercial vendors the cost can be a small fraction of primary data collection for similar data.
- **Location access:** Satellite data allows monitoring without requiring access to the program location. This can facilitate data collection without in-person visitation to hard-to-reach areas due to private property claims, conflict or other forms of insecurity, disasters, complexity of terrain, and location remoteness.
- **Spatial coverage:** The spatial resolution and extent of satellite data are far beyond what can be achieved by many other data collection systems.
- **Coverage over time:** Satellite data can often provide the ability to analyze long-

term trends going back many decades before a program took place. Access to this data may allow researchers to reconstruct baselines and trends for outcomes of interest and other variables for impact analysis. It is likely that in the future, more satellite data will be available, and that these data will be more regularly sampled, at higher-quality, and cost less than current EOS data. Annex I provides a summary of current remote sensing systems used in LTIE.

- **Ability to observe previously hard-to-observe or unobservable key outcomes (radiometric domain):** The visible properties of remote sensing data are far beyond what can be achieved by many other data collection systems and what the human eye can observe. For example, some systems can penetrate clear water up to 50 meters deep, and extract ground features through forest canopy. They can look at changes in water quality and salinity, aquifer water levels, or forest health.

Challenges to Satellite Data

LTIEs using satellite data need to take into consideration the conditions inherent to the time the data were originally captured in their analysis. Satellite data is highly influenced by seasons and other cyclical climate patterns, such as oscillations; as well as other environmental conditions, such as drought, cloud cover, time of observation, solar conditions, and atmospheric conditions. A measure taken one day by a satellite may be non-comparable with a measure taken the next day with the same system. Indeed, much of the science of the remote science community is targeted towards ensuring cross-comparability of remotely sensed measurements across time, space, and instruments. Without substantial knowledge and technical ability, evaluations cannot make accurate comparisons.

Box 10: Predicting Poverty

Satellite imagery can be used to estimate more than just agricultural yields. For some time, researchers have used satellite imagery taken at night to estimate levels of economic activity. Marshall Burke and David Lobell have updated this technique combining additional details found in daytime satellite imagery to improve remote sensing predictions of income and wealth in developing countries (Jean et al., 2016). Such estimates will provide an additional data source for long-term impact evaluations focused on aggregate geospatial outcomes, especially in countries that lack recent census data and where surveys are not easy to undertake.

Although satellite instruments are often thought of as neutral observers, there can be bias at the instrument level which may be caused by the intended use of a system. Indeed, much of the use of satellite systems now actually extends beyond the original design of the system. There is

often a lack of ground-control or other validation for these new uses. Satellite systems almost always measure a proxy of the desired variable and not the variable itself. For example, economic development is often estimated by economists through increases in the amount of light received at the sensor but not omitted by the sun (nighttime lights). Such proxy measures are strongly correlated to the measure of interest, but they are not the measure of interest per se.

Geographic Information System (GIS)

Geographic information system (GIS) data can measure changes in outcomes over time as well as identify the location and extent of an impact. To use GIS data in a LTIE, it is essential to have a geographic unit where interventions and the results of the interventions can be measured. These units can be households, communities, or even agricultural fields. Evaluations that are primarily GIS-based have been focused on natural resources, but have the potential to expand beyond this application to infrastructure and other socioeconomic programs. The lack of non-natural resource applications is likely due to in part the remote sensing expertise occurring mostly within the geophysical community although this is changing as more remotely sensed products are delivered ready to use to the wider research and non-scientific community.

Mobile Data

As the availability of cellular phone coverage expands, mobile penetration rates continue to rise in developing countries. Along with this expansion, information gathered through mobile phones is becoming an increasingly valuable source of big data. The term "big data" refers to exceptionally large quantities of data that can be used to extract patterns and learnings using machine-learning techniques to shed light on activity and project outcomes in the short and long-term.

Telecommunications companies keep records of large data sets detailing the number, frequency, location, and length of calls. Often, companies collect this data on a daily basis and routinely wipe it without a profitable incentive to utilize the data. By collaborating with cellular providers in emerging markets, implementers and evaluators have the potential to use such data sets to conduct impact evaluations that estimate outcomes such as wealth, migration patterns, social networks, and financial behavior.

Long-term impact evaluations can utilize mobile data to measure key economic development outcomes. Mobile data can be particularly useful in tracking geographic and migratory trends through analyzing call detail records. Moreover, mobile data can be used to measure outcomes such as individual wealth (see Box 11 on the next page) and could provide a cost-effective way of assessing such long-term outcomes. For example, if researchers have original phone numbers and records from beneficiaries, they can follow up with those beneficiaries to perform long-term evaluations without conducting costly field surveys. Indeed they could substitute such surveys for IVR/SMS surveys combined with utilizing mobile data records to assess long-term income impact. However, it should be noted that many developing country consumers have

Box 11: An Example of Uses of Mobile Data

One of the most prolific forms of big data from mobile phones is call detail records. Blumenstock, Cadamuro, & On (2015) demonstrated that analyzing call detail records could actually be a predictor of wealth in places like Rwanda, and later Afghanistan. Call detail records, as well as other mobile data, can be useful for understanding people's lives as well as mapping their activity and migratory patterns over time through data visualization (Blumenstock, 2011). Mobile data is particularly well-suited for generating time series data sets and, as Blumenstock, Cadamuro, & On (2015) demonstrated, can be used to track long-term outcomes, such as wealth.

multiple phones or SIM cards, and often change carriers frequently, so tracing such data trails over time can be challenging. If accessible, mobile data can prove to be an invaluable source of information about beneficiaries.

Challenges to Using Mobile Data

There are significant barriers to utilizing this kind of data for long-term evaluations. In order to access these data sets, evaluators must establish partnerships with telecommunications networks, which often take the form of lengthy negotiations. Often this process may require Non-Disclosure Agreements (NDAs) and IRB approval, and research teams should budget the necessary time to get these agreements in place before embarking on mobile data analysis. If access to such datasets is granted, research teams still need highly trained technical experts and advanced machine-learning techniques to analyze the data and draw conclusions about the subjects. Programs interested in using mobile data in the context of an LTIE may consider forming academic and private partnerships.

VI. RECAP OF DECISION-MAKING FOR LTIE DESIGNS

In this guide, we have discussed methods for evaluating long-term impact, data sources, and different scenarios for setting up LTIEs. As program managers consider whether or not to invest in an LTIE, they may consult Figure 3 on the next page where we list some of the key questions to answer in the decision-making process.

Selecting the Method

Selecting the most appropriate method for an LTIE can take time and involves working closely with evaluators to determine the feasibility of various methods and available data sources. Figure 4 on page 33 summarizes the conditions under which impact evaluations could be used to measure long-term impact.

Figure 3. Planning an LTIE: A Decision Tree

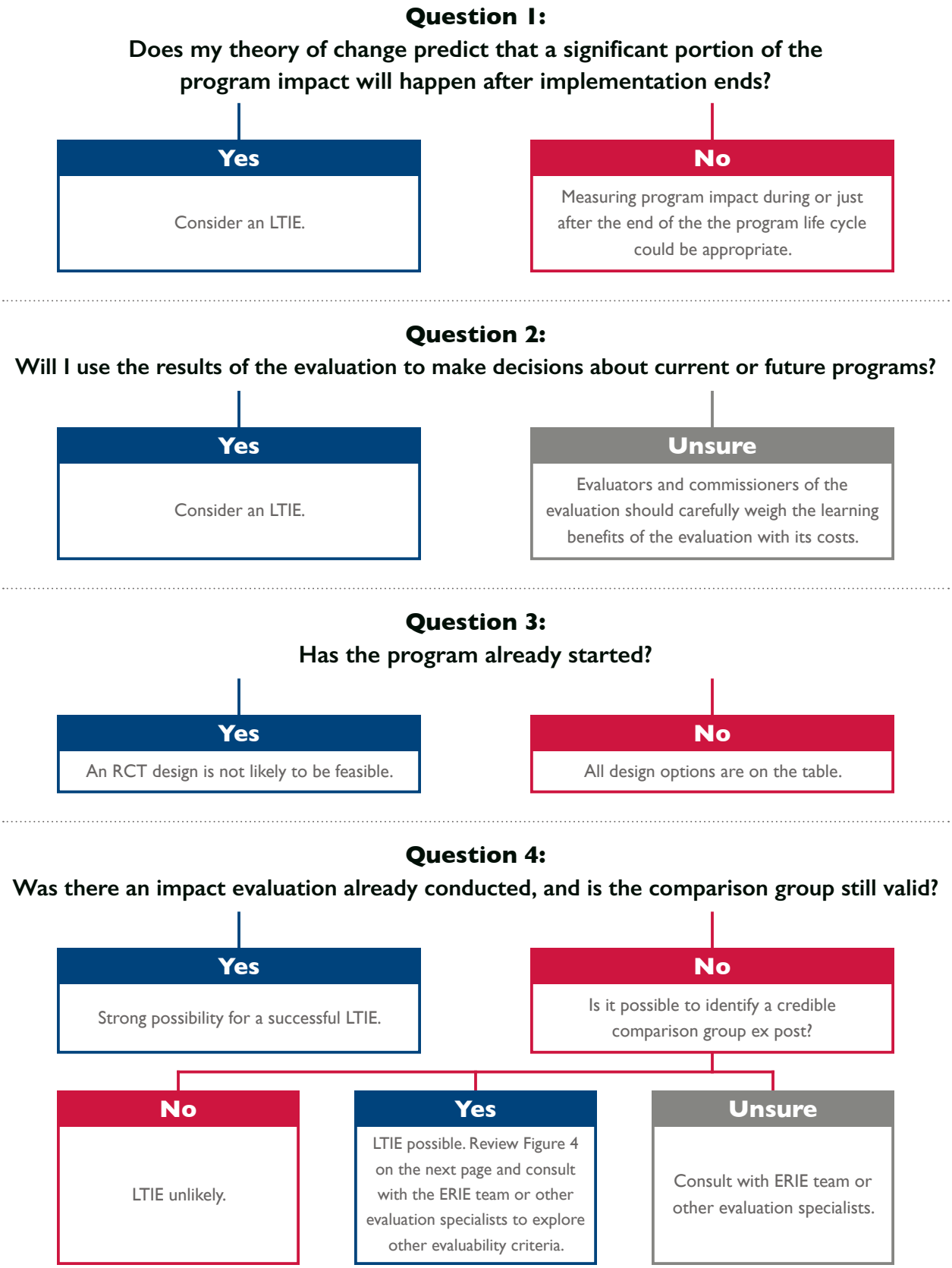


Figure 4. Selecting an Impact Evaluation Method

Program Characteristics	Potential Method
• The program has not yet started implementation. • It is possible to randomly assign an offer of participation in the program in order to create treatment and comparison groups. Or it is possible to randomly encourage eligible participants to enroll in the program. • It is possible to track outcomes for these units over the long-term.	**Randomized Controlled Trial**
• Data are available for the participating units and other units which did not participate in the program, on both the outcome of interest, and other important variables which may influence the likelihood of participation in the program. • This data are available or can be collected for at least two time periods: prior to the program's implementation, and after the implementation is complete.	**Matching**
• There is an exogenous factor which affects exposure to or participation in the program, but does not affect the outcome of interest except through the program. • Data is available on both participating units and other units who were not exposed to or did not participate in the program.	**Instrumental Variables**
• Participation in the program is determined by a strict, identifiable cut-off in a continuous variable, such as a score on an exam; age or income of participants; or distance from a certain location. • There are a sufficient number of individuals or units close to the cutoff for the evaluation sample.	**Regression Discontinuity**
• Data is available for the participating units on both the outcome of interest, and other important variables which may affect the outcome. • This data is available for (at minimum) three time periods: two periods prior to the program's implementation, and after the implementation is complete.	**Interrupted Time Series**

VII. DISSEMINATING RESULTS

Planning for dissemination of the evaluation results starts in the evaluation design phase and continues throughout its life cycle. Conducting an impact evaluation, whether long-term or not, also necessitates a larger understanding of the context for the research, the demand for the evaluation and most importantly, how the evaluators can continually engage stakeholders, disseminate the results, and ensure the research is utilized by communities and decision-makers.

Pre-Evaluation

Review the evidence. Before embarking on an impact evaluation, evaluators and implementers must review the relevant literature and summarize the evidence for the program, what gaps might remain in our understanding or program effectiveness, and how the impact evaluation will add to the wider body of evidence on the program or sector.

Plan to engage with stakeholders. As a next step, the evaluation decision-makers should develop a stakeholder engagement plan before starting the evaluation. This plan identifies the relevant individuals and groups that have an interest in the outcome of the evaluation, the likely means of engaging with the stakeholders, how the evaluation will remain accountable to beneficiaries, including beneficiaries, and how results will be shared throughout the evaluation timeline. Policy makers should be brought into the evaluation process early on and appropriately briefed after the completion of the evaluation. In addition to immediate stakeholders, evaluators can influence policy through engaging with governmental actors from civil servants to elected officials at the local, provincial and national level. Gaining government buy-in for the project and its evaluation at the onset is critical to officials making decisions using data from the intervention. Note that research teams should also make plans to manage personnel turnover as primary decision-makers will likely change over a long-term period.

Prepare for using evaluation results. Holding workshops with stakeholders offers a prime opportunity for beneficiaries, implementers, community leaders and public and private partners to discuss their priorities and concerns about the evaluation. Evaluators should note key interests and concerns of the stakeholders and consider ways of incorporating them in evaluation design and responsible results dissemination. The stakeholder workshops also allow evaluators to help decision-makers prepare for the different possible outcomes of the study, and explore the potential actions they may take depending on the evaluation results.

Complete a pre-analysis or evaluation design plan. These plans will clarify data sources, samples, and analysis methods and help disseminate the evaluation to research and academic audiences who can confirm the credibility and replicability of the results. Research teams should also consider open data standards and prepare to adhere to the standards.

Evaluation

Develop a results dissemination plan that identifies the key stakeholders and decision-

makers, and chart a course to continually keep them updated on the intervention and evaluation as well as the final results[12]. Plans would identify the answers to the following questions:

- With **whom** do we need to share results?
- What **language** should be used when sharing results with different sets of stakeholders: technical, academic, policy vernacular, or non-professional terms?
- Outside of immediate stakeholders, **who could benefit** from knowing the results? Donor community? Other comparable markets? Private industry folks?
- What vehicles can best **communicate the results** to immediate stakeholders and outside groups? Options include academic journals, policy blogs, websites, memorandum, press, webinars, published reports, conferences, workshops, and public dissemination events.

Report and communicate with evaluation audiences. Evaluation reporting varies depending on the type of evaluation and its design. Common reporting includes:

- **Evaluation Design Report:** The evaluation design report describes in detail the specific research questions the evaluation will address, relevant background literature, the analysis methods, data sources, timeline, and the plans for results dissemination. Developing the design report helps set expectations for the evaluation and come to agreement on the design, roles and responsibilities, and resources required.
- **Baseline Report:** Baseline reports describe the baseline data collection for the study. In the case of an LTIE for a new program. The baseline will help determine if the evaluation design will be valid in practice and may provide helpful information that may guide the implementation of the intervention (Gertler, Martinez, Premand, Rawlings, & Vermeersch, 2011). A baseline report will not be necessary for an LTIE that follows up an existing design or an ex-post LTIE design.
- **Follow-Up Data Collection Reports:** Some LTIEs may include several rounds of data collection. The evaluation may entail producing a report for each data collection round depending on the frequency and the information needs of evaluation decision-makers. LTIEs that rely completely on administrative data or other secondary data sources will not involve a separate data collection effort and will not produce baseline or other data collection reports.
- **Impact Evaluation Report:** In the case of an evaluation with primary data collection, the final evaluation summarizes data collection for all rounds and analyzes program results. If the evaluation is using secondary data sources, the impact evaluation report summarizes the data used and the analysis.

[12] ADS 201 provides useful guidance on creation and execution of an evaluation dissemination plan, which can be found here: https://www.usaid.gov/sites/default/files/documents/1868/201saj.pdf

- **Policy Briefs:** Policy briefs are a critical deliverable because they summarize the evaluation results for non-technical audiences and describe the main policy implications of the findings. Briefs are often short memos that can recommend specific policy prescriptions based on the results of the evaluation. Many more people will read the evaluation brief than the final report. Examples of impact evaluation policy briefs can be found on the websites of 3ie, the World Bank, or J-PAL.

Disseminating and Acting on the Results

Evaluations support poverty reduction and development when decisions-makers choose to apply the results. Policy briefs, media coverage, and direct or indirect engagement can help influence whether or not they act evaluation findings. There are also a number of other ways to disseminate. For example stakeholder workshops like the one conducted before the baseline disseminate results and provoke thoughtful discussion.

Perhaps the most critical group involved in the evaluation dissemination are the subjects themselves. Their input must be valued and promoted. Involving stakeholders of evaluations whose research team did not directly interact with subjects may be more difficult than evaluations with survey data, but it is still possible with locally relevant outreach efforts.

Finally, within the academic, donor, and implementer communities, peer-reviewed publications and conferences provide a medium by which to disseminate the evaluation results. It should also be noted that academic publications and policy briefs often influence the donor community, their priorities, expectations and request for proposal development. Publishing results and making the data available on open data platforms such as Dataverse or Figshare as well as publishing papers in open access journals increase the probability of others working in the field to learn about your evaluation and more importantly, learn *from* your evaluation.

Conclusion

The members of the ERIE team hope that this broad overview of long-term impact evaluation helps program managers within USAID and other development institutions consider how they would measure the long-term impacts of their program and apply the results. Moving forward with the actual planning and development of an LTIE will likely require input from research specialists. For more information on the ERIE mechanism and long-term impact evaluation, please email merlin@usaid.gov.

VIII. SECTOR APPLICATIONS OF LTIEs

Long-Run Impact Evaluation of the Burkinabe Response to Improve Girls' Chances to Succeed (BRIGHT) Program

In 2005, the government of Burkina Faso, with funding from the Millennium Challenge Corporation (MCC), launched Burkinabe Response to Improve Girls' Chances to Succeed (BRIGHT) to improve educational outcomes for children, especially girls. The program constructed new primary schools in rural villages with low girls' school enrollment rates and delivered other interventions designed to promote enrollment and attendance.

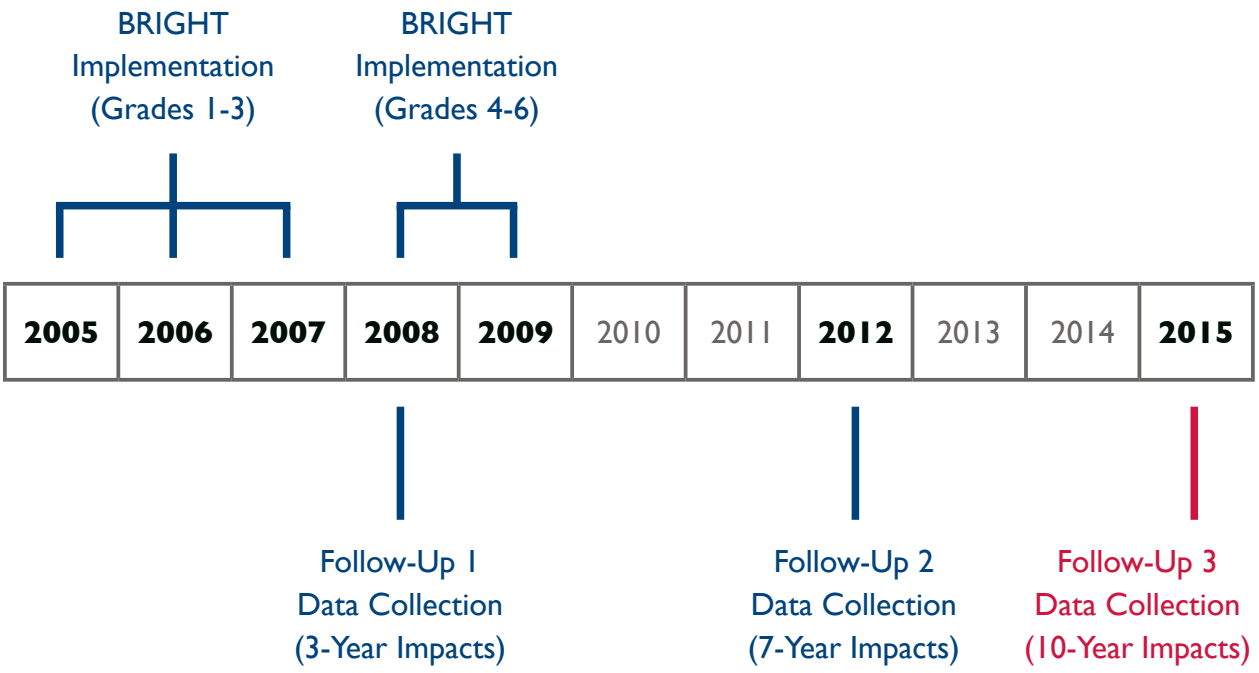
Methodology

The evaluation used a regression discontinuity design, relying on the village application scoring process and cut-off point for project eligibility to select treatment and comparison communities.

MCC followed up on an existing evaluation of a school construction program, collecting household survey data in treatment and comparison communities seven and ten years after the program started. The evaluation found that improvements in school enrollment, attendance, and test scores persisted after the program ended.

Sector	Evaluation Timeline	Method	Data Source(s)	Type
Education	2005-2015	Regression discontinuity	Household surveys and academic exams	Follow-up on existing design

Figure 5. BRIGHT Program Timeline



Findings

After three years, treatment villages saw significant improvements in enrollment and test scores. Given these results, the program was scaled-up to reach three additional grade levels in the same villages from 2008-2009. The evaluation then conducted a seven-year follow-up and found that student enrollment, attendance and test score improvements continued with additional positive impacts on child labor. At the ten-year follow-up improvements in student enrollment, attendance and test score improvements persisted although the differences between the treatment and comparison villages was slightly lower. The follow-up also found positive impacts on primary school completion rates and reductions in marriage rates for girls.

Key Takeaways and Lessons Learned

Making use of an existing evaluation design, the long-term impact evaluation successfully captured project outcomes over time, and addressed the challenge of following up with respondents over a long period of time, by surveying a new sample of households within the treatment and comparison groups during each survey round. The ten-year follow-up enabled researchers to conduct a cost-benefit assessment with longer term data showing that this relatively expensive intervention (compared to other education programs) can be cost effective.

The full evaluation report for this study can be found here.

Labor Market Returns to Early Child Development

Researchers followed up on an early childhood development study in Jamaica 20 years later by locating and conducting surveys with most of the original study participants. This follow up study found that the intervention, which provided psychosocial stimulation to stunted toddlers, had large impacts on cognitive development and increased earnings later in life by 25 percent.

Many research questions about childhood interventions necessarily require long-term evaluation. A 1991 paper that evaluated psychological stimulation and a nutritional intervention for toddlers with stunted growth in Jamaica provided an opportunity for CEGA affiliate Paul Gertler and his team to perform a long-term follow-up on an earlier randomized control trial (Gertler et al., 2014). The stimulation intervention consisted of health workers periodically visiting families, teaching skills, and encouraging mothers to play with their children, while the nutrition intervention consisted of a milk-based nutritional supplement. These interventions were delivered for a period of two years from 1986-1988. The original experiment was designed to answer whether nutrition and stimulation interventions could have a beneficial effect on mental development in growth-stunted toddlers, and found that both the psychosocial stimulation and the nutrition intervention had positive effects on mental development, and that the two interventions had a large additive effect when provided in conjunction (Grantham-McGregor, Powell, Walker, & Himes, 1991).

Methodology

The research team was able to find and interview 105 of the children included in the experiment, which was 83% of the original sample. By interviewing these individuals—who were now in their early 20s—about their work and wage history 20 years after the original experiment, the team was able to evaluate whether these interventions affected labor market outcomes. The original study also surveyed a control group of non-stunted, non-treated toddlers, none of whom received the interventions. The team also found and interviewed this external control group, allowing the long-term comparison not only between treated and non-treated toddlers with growth stunting, but also to determine whether the interventions helped the growth-stunted children to catch up to their non-stunted counterparts in the labor market.

Sector	Evaluation Timeline	Method	Data Source(s)	Type
Early childhood development	1986-2008	Randomized controlled trial	Individual surveys and national surveys	Follow-up on existing design

Findings

The research team estimated that the psychological stimulation intervention increased individual's average earnings by 42%. The treated group had improved educational outcomes, and the research team found that stunted children who received stimulation when they were young eventually caught up to the earnings of the non-stunted comparison group. However, they did not find any long-term significant outcomes from the nutritional intervention. These divergent results between the two interventions demonstrate how long-term evaluations of childhood programs can find significant long-term outcomes that may not correlate with impacts found in the original short-term evaluation.

This full study can be found here.

Long-Run Impact Evaluation of Indigenous Land Rights in Southern Ecuador

Researchers designed a study and evaluated the impact of a land titling and land management program on deforestation five years after its completion. Using a matching approach and satellite imagery along with other primary and secondary data, researchers identified a set of comparison land plots similar to those that participated in the program. The study found no impacts of the program on deforestation.

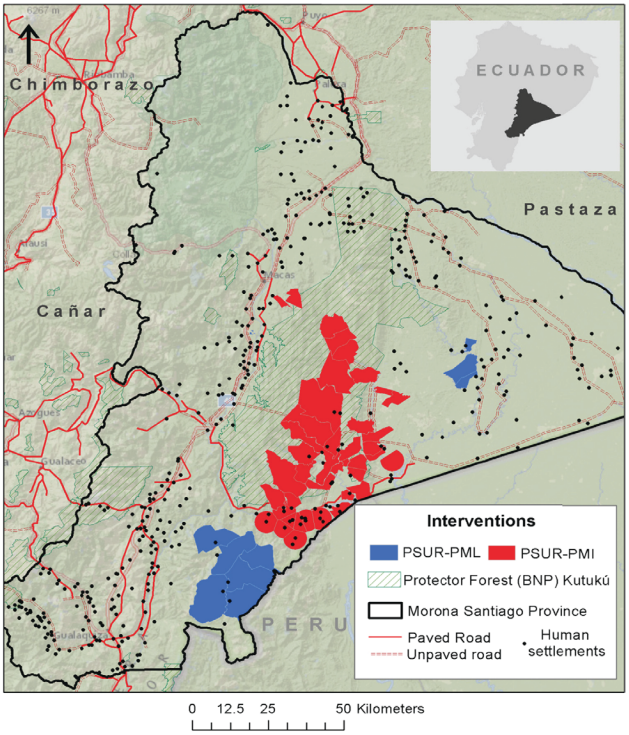
Implemented by Cooperative for Assistance and Relief Everywhere (CARE) International in collaboration with various local organizations, the Programa de Sostenibilidad y Unión Regional Sur (PSUR) aimed to empower local native Shuar population through increased property rights and to protect highly biodiverse ecosystems. From 2002-2007, the project was implemented in 52 rural communities throughout the Shuar indigenous territory in Morona-Santiago province, Ecuador that were proximate, or overlapped, with the Kutukú range, a highly biodiverse hotspot of tropical rainforest (see Figure 6 below). In the short term, the goal of the indigenous rights component was to formalize land rights, and in the medium term was to minimize land conflict and improve management practices. In the long term, the expectation was to reduce deforestation in the Kutukú area and to improve Shuar socio-economic indicators.

Land titling activities were performed in two rounds: 2002 in 9 communities (in blue), and 2003 to 2007 in 43 communities (in red), which also received additional enhanced land management plans and training.

Timeline

- PSUR was first implemented in the 9 communities from 2002 to 2003 and included land titling and a mandatory basic land management plan.
- PSUR was fully implement from 2003 to 2007. It included the same components as the first phase plus an enhanced management plan, which included natural resource management and training activities.
- PSUR was evaluated in 2007 by CARE after the program ended.
- The land titling component of PSUR

Figure 6. Morona-Santiago Province Intervention Area



The study area and locations of titling and management planning treatments. (For interpretation of the references to color in this Figure 6 legend, the reader is referred to the web version of this article.)

Sector	Evaluation Timeline	Method	Data Source(s)	Type
Environment and land tenure	2002-2012	Non-parametric matching, GEOIE	Satellite imagery and other geospatial data	Ex-post, long-term impact evaluation

is evaluated by GIE (Buntaine, Hamilton, & Millones, 2015).

Methodology

The evaluation used a quasi-experimental matching design to measure impact after the program had ended. Treatment plots located in the PSUR intervened areas were matched with comparison plots of similar social and environmental characteristics located outside of PSUR area. The goal was to examine whether areas intervened by PSUR showed less forest loss than areas where the program was not present.

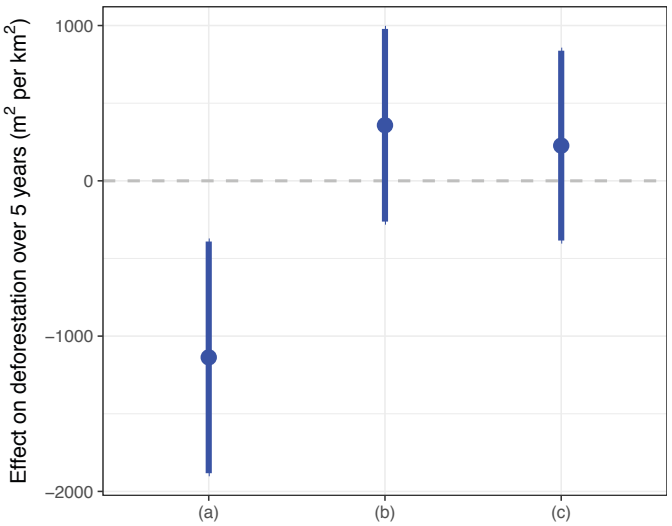
Data on forest loss was collected from 2000 to 2012 as part of the Global Forest Cover (GFC) Landsat derived product (Hansen et al., 2013). Researchers collected information on the intervention areas locations and extent with the aid of local organizations, and used treatment and comparison data from various remotely sensed and GIS data sources.

Findings

Researchers found no evidence of impact in deforestation in the five years following legal recognition of lands promoted by PSUR (see Figure 7 to the left). This finding applies to both rounds of land titling in all 52 communities.

There are several potential reasons for this outcome. Reducing forest loss might not be the best proxy indicator for the success of the intervention. The evaluation time period of five years from the end of the program might not be long enough for the effect to manifest itself. It is possible that deforestation was less likely in the areas evaluated, either

Figure 7. Five-Year Deforestation Tracking



Difference in differences over five years for PSUR plots with legalization plan (PML) and title versus non-PSUR plots with no plan or title, 2002–2012. Notes: Figure 7 shows treatment effect of tenure status for models as follows: (a) covariates, no pre-matching; (b) no covariates, pre-matching; (c) covariates, pre-matching.

because of the remoteness of the region, the land use practices, the institutional arrangements of the Shuar. However, the findings are in line with current meta analyses on the effectiveness of community land titling as an incentive to reduce deforestation (Robinson, Holland, & Naughton-Treves, 2014).

The study, with full methods, results, and supplementary material can be found here.

All-Cause Mortality Reductions from Measles Catch-Up Campaigns in Africa

Data from Demographic and Health Surveys (DHS) in 25 sub-Saharan African countries allowed researchers to study the impacts of an international effort to reduce deaths from measles through mass immunization campaigns. Using variations in the timing of the campaigns across countries and regions, the study found significant impacts on child survival ten years after the Measles and Rubella Initiative (M&RI) began.

As recently as 1999, approximately 500,000 children died of measles each year in sub-Saharan Africa. The M&RI was launched in 2001 by the American Red Cross, United Nations Foundation, U.S. Centers for Disease Control and Prevention, UNICEF, and WHO. The M&RI has run a series of large-scale vaccination campaigns, often immunizing more than 90% of the targeted child population in an entire country in the matter of 1-2 weeks. By 2010, the M&RI had coordinated Supplemental Immunization Activities (SIAs) in 31 African countries, with measles incidence dropping by as much as 90%.

Sector	Evaluation Timeline	Method	Data Source(s)	Type
Health	1997-2006	Quasi-experimental design	National surveys and administrative data	Ex-post, long-term impact evaluation

Methodology

To evaluate whether the campaign was effective in reducing child mortality[13], researchers compared changes across cohorts in 25 African countries whose campaign timing varied between 2001 and 2006. They estimated both a child-level regression model that assessed the effect of the campaign on children's survival until age five, and a duration model to test whether changes in mortality rates occur contemporaneously with M&RI campaigns. This design allowed the researchers to control for country-specific trends in child mortality and other potential confounds, as well seasonal cyclicality in measles epidemics.

The continent-wide coverage of the DHS played a key role in enabling the researchers to measure child mortality using maternal interviews that provide birth and death histories.

[13] Because competing and complementary risks can mask or heighten the improvements in child survival as a result of the initiative, researchers investigated the impacts on all-cause child mortality.

Researchers used these surveys, administered 5-7 years after the campaigns for most countries, to assess the impact of "follow-up" campaigns on child survival. In addition, the American Red Cross also provided administrative data on campaign timing and coverage.

Findings

The study found large gains in survival due to the campaigns: the probability of a child's survival to five years of age improved by approximately 2.4 percentage points (reducing baseline mortality by 14%). The campaigns cost approximately $109 per child life saved, which is extraordinarily low in both absolute terms and relative to other interventions to reduce global child mortality.

Key Takeaways and Lessons Learned

The study highlights the potential for evaluating broad, multi-country, or regional programs by leveraging standardized data collection and variation in program roll-out over these sites. Moreover, the maternal birth histories elicited through DHS allow for the recovery of a child-level history that would otherwise be difficult or expensive to obtain. The ongoing data collection efforts of the DHS program will allow for even longer-term follow-up studies to assess impacts on fertility, education, and other downstream outcomes (BenYishay & Hayek, 2017). At the same time, however, the study's use of secondary data from DHS relied on broad treatment coverage of the M&RI to ensure follow-up data from treated children; other, smaller scale interventions may see less overlap with DHS (or other secondary) samples.

This full study can be found here.

Post-Project Sustainability Study of Child Health Opportunities Integrated with Community Empowerment (CHOICE) Project, Indonesia

Researchers designed and conducted a post-project impact evaluation of a child health program in Indonesia to measure whether the project had lasting impacts on health. Using a cohort matching design and a combination of new survey data and a project evaluation, researchers compared outcomes for villages that participated in the program with villages that did not. The study found positive lasting impacts on many of the health outcomes for participant villages.

NDIGD worked with Project Concern International (PCI) to conduct a Post-Project Sustainability Study (PSS) of a USAID-funded Child Health Opportunities Integrated with Community Empowerment (CHOICE) program, implemented in Indonesia between 2003 and 2007, to determine whether project interventions led to lasting improvements in infant health, nutrition, and cognitive development. The PSS assessed mother-infant pairs, seven years after cessation of funding.

CHOICE was designed to improve the health and nutrition status among children under the age of five, as well as the health status of pregnant and lactating women and mothers or caretakers of children under the age of five. CHOICE was implemented at community health centers by staff working in conjunction with village leaders and community members. Under CHOICE, volunteers provided growth monitoring, family planning, referral for curative care, immunizations, and nutrition surveillance services. This project reached 30 villages in five of the neediest sub-districts (Patia, Angsana, Pagelaran, Saketi, and Sukaresmi) in the Pandeglang District.

Methodology

To capture the long-term impact of the CHOICE project, the study followed a multi-cohort approach—children who might have been affected by the project as infants during its

A caretaker works with two new parents to measure their baby in the Pandeglang District of Indonesia.

Sector	Evaluation Timeline	Method	Data Source(s)	Type
Health	2003-2014	Ex-post, cohort study	Surveys, cognitive development test, anthropometric measures	Post-project sustainability study

implementation and who were currently aged eight to ten years old, and infants up to two years old. The data from children approximated long-term effects of the program, and the data from infants confirmed whether the knowledge and practices had continued with another cohort of children after PCI left.

Sustainability was measured in two ways: through a comparison of data collected by NDIGD in 2014 with PCI's final evaluation of CHOICE conducted in 2007; and through a comparison of 2014 data collected from the CHOICE intervention villages against a comparison group consisting of nearby villages that had not been part of the CHOICE program.

The sample population consisted of 1,947 mother-child pairs from 56 villages. Infants (under two years old) were randomly selected from current volunteers' records. In this study, sustained behavior and health status were observed indirectly through the use of a mixed-methods (both quantitative and qualitative) primary data collection.

Findings

PSS results showed a positive lasting effect in many of the indicators in CHOICE villages, for example, births attended by skilled personnel, infants' and mothers' nutrition, treatment of diarrhea, and mothers receiving at least two tetanus toxoid injections. However, other project results, such as infant immunization, Vitamin A, and treating water effectively, were not sustained.

Several differences between intervention and comparison villages were found to be statistically significant in 2014. Positive differences were observed in terms of the percent of infants put to the breast within one hour of birth, percent who had anything by bottle in the last 24 hours, and percent of mothers that sought outside advice or treatment for their infants with diarrhea.

The qualitative analysis in general shows that CHOICE had a long-lasting positive effect on the people who were part of the program. Those interviewed reported that the CHOICE program had changed the minds of mothers with respect to childbirth and infant health, and had paved the way to a continued improvement in sanitation facilities.

More information on the study can be found here.

Environment: Indigenous Land Rights and Deforestation: Evidence from the Brazilian Amazon

Researchers designed a study and evaluated the impact of a land rights formalization program on deforestation several years after rights were formalized. Using a propensity score matching approach and satellite imagery, researchers identified a set of comparison land plots similar to those that participated in the program. The study found no long-term impacts on deforestation.

Indigenous communities in Brazil suffered from weak protection of their communal lands for many years; the country's 1988 constitution attempted to address this by granting communities important protections. The Brazil Indigenous Lands Project (PPTAL), a collaboration between the World Bank and German KfW, aimed to put these rights into force by supporting the demarcation and formalization of more than 38 million hectares of largely forested lands between 1995 and 2008. Researchers studied the program's impact using a 30-year time-series of satellite-based forest cover data.

Methodology

The evaluation involved two complementary methodologies: The researchers used propensity score matching on the basis of pre-program levels and trends in deforestation and covariates, such as land area, population, slope, elevation, and distance to the closest river and road. The second method employed grid cell-level fixed effects to adjust for the non-random demarcation of indigenous communities.

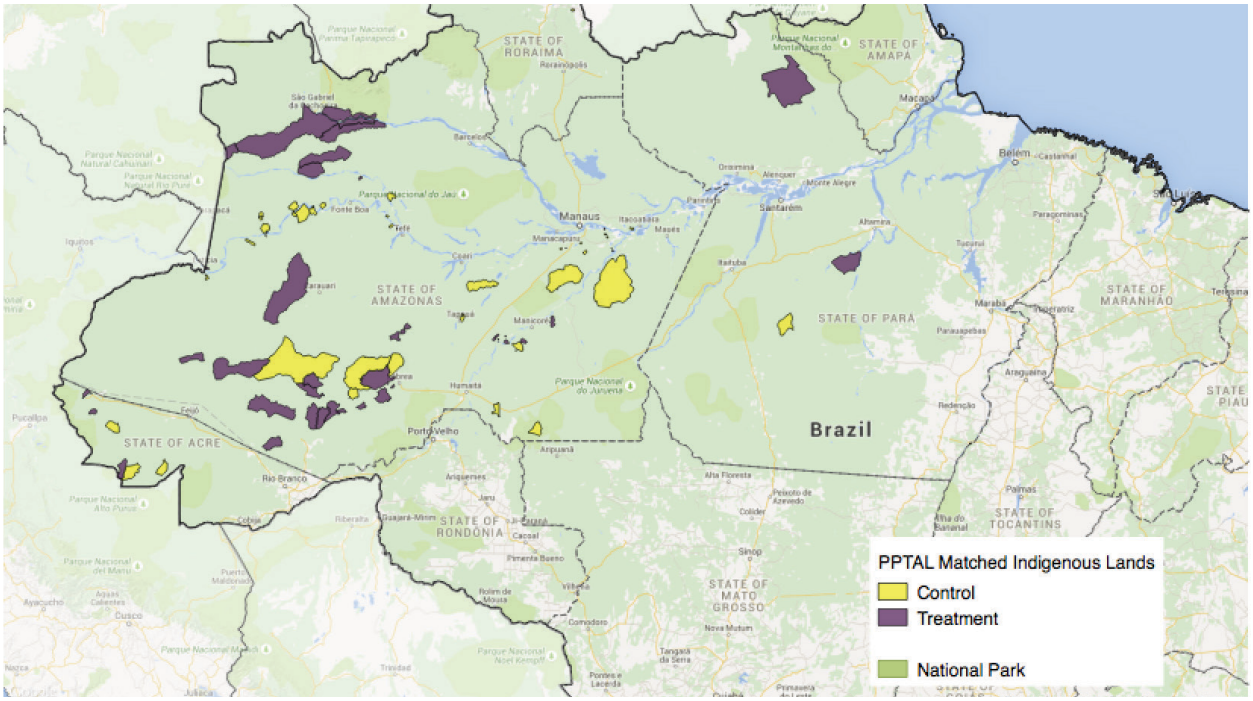
The evaluation merged the NASA data with administrative data from PPTAL's implementing agency FUNAI on the month and year in which each of the 151 communities were initially studied, as well as the dates at which communities completed the demarcation and approval stages. FUNAI also provided the geographic boundaries of each communities.

Sector	Evaluation Timeline	Method	Data Source(s)	Type
Environment and land rights	1982-2014	Propensity score matching and fixed effects	Satellite data and administrative dats	Ex-post, long-term impact evaluation

Findings

The researchers found no differences in deforestation dynamics that could be attributed to the PPTAL project.

Figure 8. Brazil Indigenous Lands Project (PPTAL) Project Map



The project exceeded its short-term goals by identifying and delimiting boundaries for more than 100 communities. In the medium term, the project was also successful in supporting the maintenance of these boundaries, GPS training, surveillance routines, transportation acquisition, and establishment of comparison posts. However, the project's impacts on deforestation had yet to be assessed. The researchers used follow-up data through 2014 using the NASA Land Long-term Data Record NDVI measures of vegetative greenness, meaning most lands were observed for upwards of a decade following their demarcation. The research now found no effect of the PPTAL's protections on satellite-based greenness measures, even for communities that received support for surveillance and enforcement of these rights. Importantly, the researchers observed low counterfactual rates of deforestation on communities' lands between 1982 and 2014, suggesting that the communities protected through PPTAL were not as threatened as initially expected. Data on forest greenness using satellite measures was not yet available in 1995 when PPTAL first identified its priority communities.

Key Lessons

The study's results contrast the program targeting possible with newly available satellite data to that possible during the program's actual design. It also suggests far less selection on outcomes (e.g., effective targeting) was done, making quasi-experimental models more likely to be valid than critics might argue.

The full study can be found here.

REFERENCES

Abadie, A., & Gardeazabal, J. (2003). The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review, 93*(1), 113–132. https://doi.org/10.1257/000282803321455188

Baird, S., Hicks, J. H., Kremer, M., & Miguel, E. (2016). Worms at Work: Long-Run Impacts of a Child Health Investment. *The Quarterly Journal of Economics, 131*(4), 1637–1680. https://doi.org/10.1093/qje/qjw022

BenYishay, A., & Hayek, D. (2017). *Child Mortality and the Quantity-Quality Trade-Off* (Abstract). College of William & Mary and the University of New South Wales. Retrieved from https://www.business.unsw.edu.au/About-Site/Schools-Site/Economics-Site/Documents/BenYishay-Hayek-child-mortality.pdf

BenYishay, A., Heuser, S., Runfola, D., & Trichler, R. (2016). *Indigenous Land Rights and Deforestation: Evidence from the Brazilian Amazon* (Working Paper). Williamsburg, VA: AidData, College of William & Mary. Retrieved from http://docs.aiddata.org/ad4/files/wps22_indigenous_land_rights_and_deforestation.pdf

BenYishay, A., & Kranker, K. (2015). All-Cause Mortality Reductions from Measles Catchup Campaigns in Africa. *Journal of Human Resources, 50*(2), 516–547. https://doi.org/10.3368/jhr.50.2.516

Blumenstock, J. (2011). *CDR-Inferred Locations of a Single Individual Over 4 Years (a40)*. University of California, Berkeley. Retrieved from https://youtu.be/DoWC85nwUeY

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting Poverty and Wealth from Mobile Phone Metadata. *Science, 350*(6264), 1073–1076. https://doi.org/10.1126/science.aac4420

Bohra-Mishra, P., Oppenheimer, M., & Hsiang, S. M. (2014). Nonlinear Permanent Migration Response to Climatic Variations but Minimal Response to Disasters. *Proceedings of the National Academy of Sciences of the United States of America, 111*(27), 9780–9785. https://doi.org/10.1073/pnas.1317166111

Buntaine, M. T., Hamilton, S. E., & Millones, M. (2015). Titling Community Land to Prevent Deforestation: An Evaluation of a Best-Case Program in Morona-Santiago, Ecuador. *Marco Millones, 33*(0), 32–43. https://doi.org/10.1016/j.gloenvcha.2015.04.001

Children's Investment Fund Foundation. (n.d.). *Rolling Out a Global Emissions Accounting System for Cities* (Web Page). Online. Retrieved from https://ciff.org/grant-portfolio/rolling-out-global-emissions-accounting-system-cities

Croke, K., Hicks, J. H., Hsu, E., Kremer, M., & Miguel, E. (2017). Should the WHO Withdraw Support for Mass Deworming? *PLOS Neglected Tropical Diseases, 11*(6). https://doi.

org/10.1371/journal.pntd.0005481

Davis, M., Ingwersen, N., Kazianga, H., Linden, L., Mamun, A., Protik, A., & Matt Sloan. (2016). *Ten-Year Impacts of Burkina Faso's BRIGHT Program* (Report). Princeton, NJ: Mathematica Policy Research. Retrieved from https://www.mathematica-mpr.com/our-publications-and-findings/publications/ten-year-impacts-of-burkina-fasos-bright-program

Garces, E., Thomas, D., & Currie, J. (2002). Longer-Term Effects of Head Start. *American Economic Review, 92*(4), 999–1012. https://doi.org/10.3386/w8054

Gathani, S., Stoelinga, D., & Santini, M. (2013). Innovative Techniques to Evaluate the Impact of Private Sector Development Reforms: An Application to Rwanda and 11 Other Countries. Presented at the MPSA Annual Conference, Midwest Political Science Association. Retrieved from https://www.researchgate.net/publication/315516764_Innovative_techniques_to_evaluate_the_Impact_of_private_sector_development_reforms_An_application_to_Rwanda_and_11_other_countries

Gertler, P., Heckman, J., Pinto, R., Zanolini, A., Vermeersch, C., Walker, S., … Grantham-McGregor, S. (2014). Labor Market Returns to an Early Childhood Stimulation Intervention in Jamaica. *Science, 344*(6187), 998–1001. https://doi.org/10.1126/science.1251178

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. J. (2011). *Impact Evaluation in Practice* (1st ed.). Inter-American Development Bank and World Bank. http://doi.org/10.1596/978-0-8213-8541-8

Grantham-McGregor, S. M., Powell, C. A., Walker, S. P., & Himes, J. H. (1991). Nutritional Supplementation, Psychosocial Stimulation, and Mental Development of Stunted Children: The Jamaican Study. *The Lancet, 338*(8758), 1–5. https://doi.org/10.1016/0140-6736(91)90001-6

Guzman, J. C., Brown, D., & Khatiwada, L. (2016). *Millennium Challenge Corporation Impact Evaluation: Water and Sanitation Component of the Ghana Compact* (Final Report). Notre Dame, IN: Notre Dame Initiative for Global Development, University of Notre Dame.

Hansen, M. C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., … J. R. G. Townshend. (2013). High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science, 342*(6160), 850–853. https://doi.org/10.1126/science.1244693

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Stefano Ermon. (2016). Combining Satellite Imagery and Machine Learning to Predict Poverty. *Science, 353*(6301), 790–794. https://doi.org/10.1126/science.aaf7894

Lipscomb, M., Mobarak, A. M., & Tania Barham. (2013). Development Effects of Electrification: Evidence from the Topographic Placement of Hydropower Plants in Brazil. *American Economic Journal: Applied Economics, 5*(2), 200–231. https://doi.org/10.1257/app.5.2.200

Miguel, E., & Kremer, M. (2004). Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica, 72*(1), 159–217. Retrieved from http://cega.berkeley.edu/assets/cega_research_projects/1/Identifying-Impacts-on-Education-and-Health-in-the-Presence-of-Treatment-Externalities.pdf

Oster, E. (2017). Unobservable Selection and Coefficient Stability: Theory and Evidence. *Journal of Business & Economic Statistics*. https://doi.org/10.1080/07350015.2016.1227711

Reinikka, R., & Svensson, J. (2010). Fighting Corruption to Improve Schooling: Evidence from a Newspaper Campaign in Uganda. *Journal of the European Economic Association, 3*(2–3), 259–267. https://doi.org/10.1162/jeea.2005.3.2-3.259

Robinson, B. E., Holland, M. B., & Naughton-Treves, L. (2014). Does Secure Land Tenure Save Forests? A Meta-Analysis of the Relationship Between Land Tenure and Tropical Deforestation. *Global Environmental Change, 29*, 281–293. https://doi.org/10.1016/j.gloenvcha.2013.05.012

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology, 66*(5), 688–701. https://doi.org/10.1037/h0037350

Seftor, N. S., Mamun, A., & Schirm, A. (2009). *The Impacts of Regular Upward Bound on Postsecondary Outcomes 7-9 Years After Scheduled High School Graduation*. Princeton, NJ: Mathematica Policy Research. Retrieved from https://www.mathematica-mpr.com/our-publications-and-findings/publications/the-impacts-of-regular-upward-bound-on-postsecondary-outcomes-79-years-after-scheduled-high-school-graduation

United States Agency for International Development. (2013). *Impact Evaluations* (Technical Note). Online. Retrieved from https://www.usaid.gov/sites/default/files/documents/1870/IE_Technical_Note_2013_0903_Final.pdf

United States Agency for International Development. (2017). *Developing a Project Logic Model (and its Associated Theory of Change)* (How-To Note). Online. Retrieved from https://usaidlearninglab.org/sites/default/files/resource/files/project_logic_model_how_to_note_final_sep1.pdf

United States Department of Education. (n.d.-a). *Early Learning: Frequently Asked Questions* (Web Page). Online. Retrieved from https://www.ed.gov/budget14/faqs/early-learning

United States Department of Education. (n.d.-b). *Upward Bound Program* (Web Page). Online. Retrieved from https://www2.ed.gov/programs/trioupbound/index.html

United States Department of Health and Human Services. (n.d.). *Code of Federal Regulations, Title 45: Protection of Human Subjects* (Web Page). Online. Retrieved from https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html

ANNEX 1: FURTHER INFORMATION SATELLITE DATA

Earth Observing System (EOS) can be divided into passive systems or active systems, with passive systems the most common and most utilized with the evaluation community.

Passive sensors use natural energy from the sun to detect and measure reflected energy from the locations being sensed. Passive sensors detect sunlight reflected from the earth's features in a wide region of the electromagnetic spectrum. Passive sensors data are widely used in remote sensing studies, and the U.S.-funded Landsat mission is the oldest running earth observation program equipped with a passive sensor. Other examples of passive sensors include MODIS (Moderate Resolution Imaging Spectroradiometer), ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer), AVHRR (Advanced Very-High Resolution Radiometer), World View, and Quick Bird. Passive sensor satellite data are used in studying agriculture, forestry, biodiversity, climate change impact, disasters, human health, land use change, and almost all human and physical components of the earth.

Active sensors use their own source of energy for illumination to capture information about the earth. These are less common than passive sensors. One of the advantages of active sensors is their ability to obtain measurement anytime regardless of daylight or weather condition. Spaceborne active sensor satellites are also useful in observation of atmospheric phenomenon as well. Examples are LIDAR (Light Detection and Ranging), SAR (Synthetic Aperture Radar), and PALSAR (Phased Array L-band Synthetic Aperture Radar). Some application of active sensors includes weather forecasting, studying soil moisture, bathymetric and topographic mapping, groundwater analysis, and snow or ice studies.

Landsat is currently in its eighth iterations (as seen in Figure 9 on the next page), with a ninth instrument planned for launch around 2020. Other countries now have similar Landsat-like systems such as the ESA SPOT system, the Japanese ASTER system, the Indian IRS system and the Chinese-Brazilian CBERS system. The benefits of the Landsat system are the regular collection system used across differing instruments, the long-term record of data now with a history of 45-years, the free availability of data, the high level of pre-processing and processing that can be achieved before the data arrives at the user, and the broad applicability of the data generated. Landsat is planned to continue well beyond 50-years, and similar sensors such as those listed above will likely be in orbit continuously into the future.

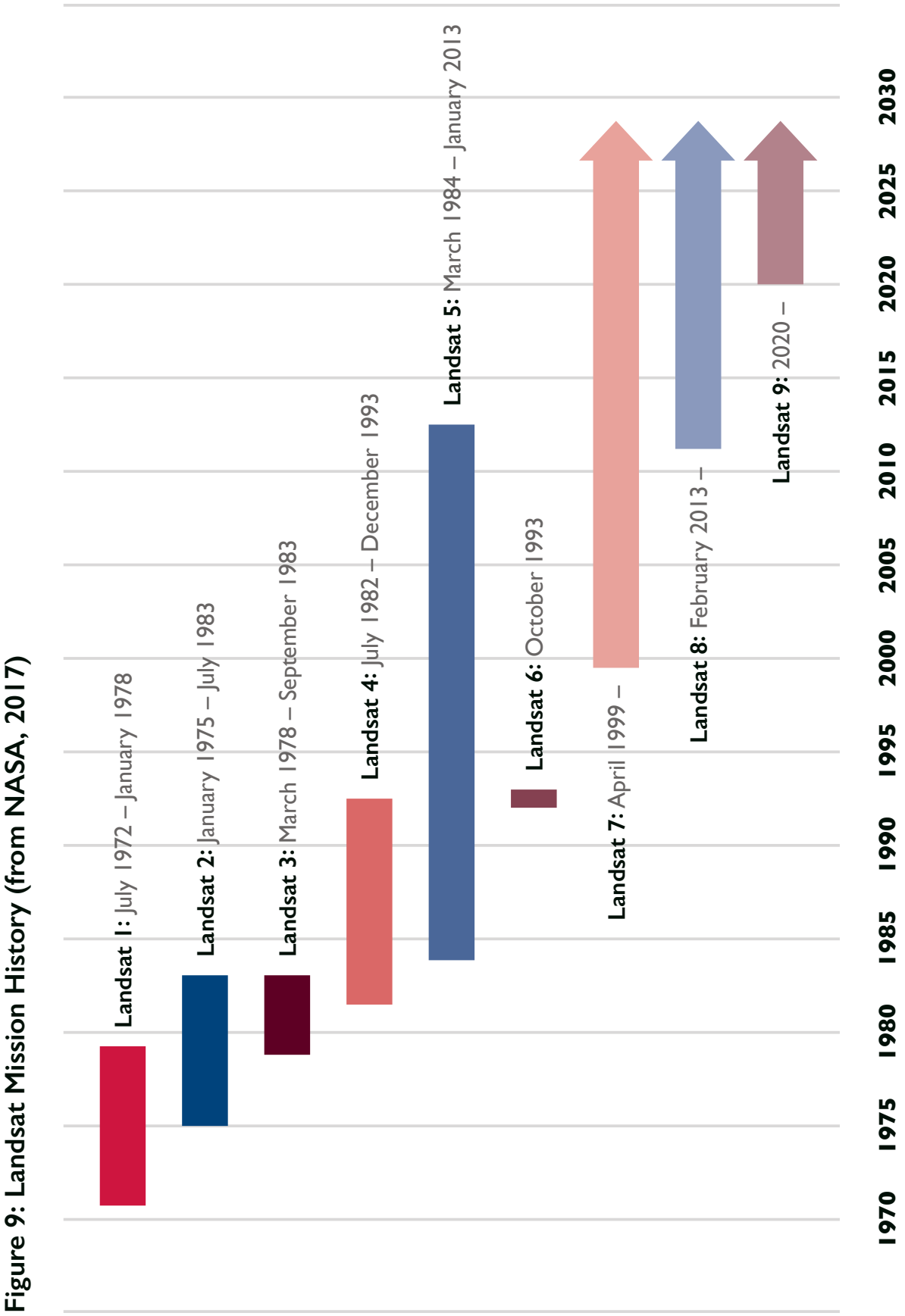


Figure 9: Landsat Mission History (from NASA, 2017)

This page is intentionally left blank.

1300 Pennsylvania Avenue NW
Washington, DC 20523

Email: MERLIN@USAID.gov
Phone: (202) 712-0000
Fax: (202) 216-3524

USAID.gov/GlobalDevLab/MERLIN/ERIE