GLOBAL PARTNERSHIP
FOR YOUTH EMPLOYMENT

Measuring Success of Youth Livelihood Interventions

A Practical Guide to Monitoring and Evaluation

Kevin Hempel
Nathan Fiala

THE WORLD BANK
The Youth Employment Network
unicef
Understanding Children's Work
An Inter-Agency Research Cooperation Project
international youth foundation

The complete guide, individual notes, and supplemental resources are available online at www.gpye.org

1818 H Street NW
Washington DC 20433
Telephone: 202-473-1000
www.worldbank.org

Suggested citation:
Hempel, Kevin, and Nathan Fiala. 2011. *Measuring Success of Youth Livelihood Interventions: A Practical Guide to Monitoring and Evaluation.* Washington, DC: Global Partnership for Youth Employment.
http://www.gpye.org/measuring-success-youth-livelihood-interventions

Cover: Kathryn Werthman, International Youth Foundation; Cover photo: Nathan Fiala
Photo on page 26: Courtesy ILO-YEN; all other interior photos courtesy IYF.

We welcome your feedback to further improve this guide.
For comments and suggestions, please contact:

Kevin Hempel
khempel@worldbank.org

Nathan Fiala
nfiala@diw.de

Measuring Success of Youth Livelihood Interventions

A Practical Guide to Monitoring and Evaluation

Kevin Hempel
Nathan Fiala

Learning is not attained by chance.
It must be sought for with ardor and attended to with diligence.

— Abigail Adams

Contents

BOXES

FIGURES

TABLES

FOREWORD

In its 2007 *World Development Report: Development and the Next Generation*, the World Bank underscored the world's historic "youth bulge" of 1.3 billion young people and the urgent need for governments and the development community to invest in this younger generation. The report focused on the pivotal stages of life that young people must navigate successfully to unleash their extraordinary potential. It also stressed the critical importance of identifying and sharing best practices, tested policies, and scalable lessons learned so that those investments can have the greatest possible impact, now and in the future.

Since that report was released, we have seen, often in dramatic ways, what happens when more and more young people do not make the transition to adulthood successfully. Rising youth unemployment, for example, has been a factor underlying much of the unrest in today's world. Instead of making progress, we seem to be losing ground. What policies, therefore, need to be in place to truly address the growing global challenges facing our young people? Given limited resources, how can we make sure we are investing in the most effective programs, especially when many have not been rigorously evaluated and tested? How do we assess what we know and what we still need to know to scale up workable solutions for youth?

In order to build and disseminate evidence around proven methods to improve youth livelihood outcomes, the World Bank established the Global Partnership for Youth Employment in 2008. The partnership comprises the Arab Urban Development Institute, the International Youth Foundation, Understanding Children's work, and the Youth Employment Network. Convinced that more and better evidence can improve the design and outcomes of future youth programs and strengthen the entire development field, we have been working together to promote impact evaluations and related learning tools and strategies.

We are therefore pleased to present this publication, *Measuring Success of Youth Livelihood Interventions: A Practical Guide to Monitoring and Evaluation*, which offers a comprehensive and accessible introduction to this important topic of monitoring and evaluation and its application in the field of youth employment and livelihood development. We hope this guide will help

practitioners make informed decisions in choosing the evaluation frameworks that can most benefit their programs and organizations and encourage greater transparency in the process. If we can draw more robust evidence from all the good work that we have accomplished over the past decade, and then share that knowledge with the broader development community, we will have a far stronger voice in convincing policymakers to adopt successful programs. Developing systematic, broad-based programs with predictable outcomes will also mobilize support for taking them to scale so we can reach far more young people in the years ahead.

We look forward to continuing our work together to provide young people with the skills and opportunities they need to be successful workers, entrepreneurs, parents, citizens, and indeed, leaders. This publication is one key effort by our Global Partnership to achieve this goal.

Arup Banerji
Director, Social Protection and Labor
The World Bank

Susana Puerto Gonzalez
Manager
Youth Employment Network

William S. Reese
President
International Youth Foundation

Furio Rosati
Director
Understanding Children's Work

ACKNOWLEDGMENTS

This guide would not have been possible without the contributions of many colleagues and friends. First and foremost, we would like to thank Mattias Lundberg and Wendy Cunningham for their overall guidance, feedback, and support throughout the planning and writing process, as well as the World Bank's Director for Social Protection and Labour, Arup Banerji.

Further, we would like to acknowledge our colleagues from the Global Partnership for Youth Employment who have made significant contributions in providing content and suggestions for parts of this manual. They include Ibrahim Al-Turki from the Arab Urban Development Institute; Awais Sufi, Travis Adkins, Dan Oliver, Jack Boyson, Susan Pezzullo, and Hannah Corey from the International Youth Foundation; Markus Pilgrim, Susana Puerto Gonzales, and Drew Gardiner from Youth Employment Network; and Gabriella Breglia from Understanding Children's Work. The writing of this guide was truly a team effort.

We are also extremely grateful to our peer reviewers Gloria La Cava from the World Bank, and Fiona Macaulay and Veronica Torres from Making Cents International.

We would like to recognize the many colleagues inside and outside the World Bank who have taken the time to read earlier drafts of this manual and to provide invaluable feedback, suggestions, and background materials. World Bank contributors include Paloma Acevedo, Juliana Arbelaez, Marinella Ariano, Carlos Asenjo Ruiz, Stefanie Brodmann, Shubha Chakravarty, Yoonyoung Cho, Viola Erdmannsdoerfer, Tanja Lohmann, Florentina Mulaj, Suleiman Namara, David Locke Newhouse, Azra Kacapor Nurkic, Yaa Oppong, Michelle Rebosio, Leopold Sarr, Haneen Ismail Sayed, Devin Silver, Stavros George Stavrou, and Cornelia Tesliuc.

We also thank Dino Linares of Colectivo Integral de Desarrollo; Pia Saunders of Education for Employment Foundation; Alexandre Kolev of ILO-ITC; David Rosas of Inter-American Development Bank; Kristin Hausotter and Bettina Silbernagl of GIZ; Anne Golla from International Centre for Research on Women; Caroline Jenner and Shannon Wendt of Junior Achievement; Whitney Harrelson of Making Cents International; Rewa Misra of Mastercard Foundation; Leah Katerberg and Scott Ruddick

of MEDA; Sonya Silva and David Woollcombe of PeaceChild International, Karen Austrian of Population Council; Rani Deshpande from Save the Children; Justin Sykes of Silatech; Nader Kabbani of the Syria Trust for Development; Rachel Blum of USAID; and Helen Gale of Youth Business International.

In addition, we would like to acknowledge Paul Gertler, Sebastian Martinez, Patrick Premand, Laura Rawlings, and Christel Vermeersch for their excellent publication *Impact Evaluation in Practice*. This guide builds on their work, adapting some of the material and illustrations to the youth livelihood field and providing a more concise presentation of impact evaluation methods. Further, we would like to thank the participants of the Youth Employment Network evaluation clinics in Nairobi, Beirut, and Geneva; the many active members of the YEN Groupsite; and the participants of the Making Cents International Global Youth Enterprise Conference and World Youth Congress 2010, whose questions and queries have all informed the content of this document.

A special thanks goes to our editor, Kris Rusch, whose advice, responsiveness, and willingness to work under a very tight timeline were absolutely invaluable. We also thank Caroline Esclapez, Gillian McCallion, Lynde Pratt, and Kathryn Werthman for their contributions to the design of this work.

Finally, Kevin Hempel would like to thank Carla Rojas for her patience.

Any errors that may remain in this document are our own.

INTRODUCTION

Programs to actively support young people's employment prospects have existed for decades in industrialized countries; however, they are relatively new in developing nations. In a broad sense, youth livelihood interventions support young people's means to earn a living, and include training, public service, youth entrepreneurship, and financial services. More narrowly, many practitioners define youth livelihood programs as activities targeting particularly vulnerable and marginalized groups in the informal economy, with a specific focus on self-employment. This guide adopts the broader definition and includes workforce development for the formal sector.

As a relatively new and innovative sector, few interventions have been rigorously evaluated. In fact, most practitioners could cite only a handful of examples. But what does *rigorous* really mean? Which methods are rigorous enough, and which ones are not? To practitioners, it may often seem obvious that our intervention is yielding the desired results. Why spend our limited resources on an expensive evaluation if we could instead use the money to provide services to more young people?

For those not directly involved in the intervention, its effectiveness is not always obvious. Policymakers and donors want credible, transparent results that satisfy some minimum standards of reliability. They are often looking for evaluations that use established social science research methods, which can provide robust estimates on how an intervention affected the typical program participant. Practitioners, in turn, though concerned with providing quality information about their programs, may feel that rigorous evaluations, with their complexity, potential costs, and other resource requirements, are often unrealistic and out of reach.

Areas of intervention for youth livelihood development programs

- *Training and skills development*
- *Subsidized employment, including wage subsidies, public works and public service programs*
- *Employment services, including job search assistance and placement support*
- *Youth enterprise and entrepreneurship*
- *Youth-inclusive financial services*
- *Non-traditional programs for excluded groups*
- *Labor market regulation affecting young people*

Sources: Betcherman et al. (2007); Cunningham, Sanchez-Puerta, and Wuermli (2010); DFID (1999).

Audience

This is an introductory guide written for practitioners with no—or very limited—knowledge about impact evaluation or quantitative research methods, but who nonetheless care about demonstrating the true results of their work. It speaks to program managers and local monitoring and evaluation (M&E) officers across all types of organizations active in the youth livelihood field: local and international NGOs, local and national government officials, and bilateral and multilateral donors.

Given the diversity of backgrounds and experiences among practitioners, it is impossible to tailor this guide to everyone equally well. However, we have tried to provide a comprehensive discussion of evaluation methods for youth livelihood interventions so that readers can identify the sections most relevant to their own interests and needs.

Objective

With this guide, we aim to equip readers with the basic set of concepts and tools needed to make informed decisions about how to best evaluate their programs. We seek to provide a clear understanding of the variety of evaluation options available and the

considerations that will allow practitioners to choose the most appropriate one based on learning objectives and operational context. Moreover, we describe how to manage an impact evaluation if it is the assessment method of choice.

Our overarching goal is to strengthen the foundation of sound programming and policymaking by increasing the number of quality evaluations in the youth livelihood field, thereby facilitating the scale-up and replication of successful interventions.

Focus of the Guide

The guide addresses the monitoring and evaluation of youth livelihood interventions, with a specific focus on impact evaluation. The terms *monitoring* and *evaluation* are often used jointly. However, they refer to activities that are quite different.

Monitoring tracks the implementation and progress of an intervention in order to support program administration. Monitoring

- involves the collection of data on specific implementation and results indicators.
- assesses compliance with work plans and budgets.
- uses information for project management and decision making.
- is ongoing.
- answers the question, "Are we doing the project right?"

Evaluation assesses the design, implementation, or results of an intervention in order to support new planning. Evaluation

- involves the collection of data on the design, implementation, and results of a project.
- looks at a project's relevance, efficiency, effectiveness, and sustainability.
- generates useful information about the impact of the intervention.
- is periodic; usually conducted annually at completion of a project, and includes follow up.
- answers the question, "Are we doing the right project?"

Ideally, both monitoring and evaluation should be integral parts of any program and should be planned at the program design stage. In fact, accurately assessing the success of an intervention may not be possible if the evaluation remains an afterthought that is given little priority until the program ends.

An *impact evaluation* is a type of evaluation that measures changes in the well-being of individuals, families, or communities attributable to a particular intervention. An impact evaluation answers the question: What would have happened to the beneficiaries if the program had not been undertaken? For example, if a recent graduate of a skills-training program finds a job, is it a direct result of the program, or would that individual have found work anyway? Comparing the outcomes experienced by participants with those experienced by a well-selected comparison group of nonparticipants makes it possible to establish causality. In other words, impact evaluations allow us to attribute any observed changes in the well-being of program beneficiaries to the effectiveness of our intervention.

Impact evaluation is one type of evaluation among several available, with its advantages and limitations. We believe that not every intervention requires an impact

[Definition]

Monitoring: A continuous process of collecting and analyzing information to see how well a project, program, or policy is being executed and performing against expected results.

Evaluation: A systematic, objective assessment of an ongoing or completed project design, implementation, and result to determine its relevance and the fulfillment of objectives, efficiency, effectiveness, impact, and sustainability.

Impact Evaluation: A special type of evaluation that assesses the changes in the wellbeing of individuals, households, or communities that can be attributed to a particular intervention.

Sources: Adapted from Gertler et al. (2011); Kusek and Rist (2004); OECD (1991).

evaluation and that evaluation should support programming, not the other way round. Any evaluation needs to fit the operational characteristics and context of the respective intervention, while being integrated in a larger framework that builds on an established monitoring system. That said, we also believe that much could be learned from using impact evaluation methodologies more frequently.

This guide differs from existing works in three major ways:

- First, we directly apply the concepts of M&E—and of impact evaluation in particular—to the youth livelihood sector. The book presents real-life examples, testimonies, indicators, and practical challenges as they relate to evaluating youth livelihood interventions.
- Second, we seek a balance between the practical toolkits that emphasize general monitoring and evaluation (e.g., Gosparini et al. 2004; Kellogg 1998; Kellogg 2004) and other publications that focus specifically on impact evaluation (e.g., Baker 2000; Duflo, Glennerster, and Kremer 2006; Gertler et al. 2011; Khandker, Koolwal, and Samad 2010; Ravallion 2008).
- Third, we explicitly target practitioners in the youth livelihoods field who do not have prior knowledge in research methods and evaluation and who demand a succinct, yet comprehensive, illustration of M&E and how it applies to their everyday work. Thus, in contrast to the publications above, this manual is designed to give a more concise and youth-specific presentation of the respective contents. For a more comprehensive introduction to the specific topic of impact evaluation and its practice in development, we encourage the reader to consult *Impact Evaluation in Practice* by Gertler and colleagues (2011).

Case Studies

Throughout the guide, we use the Northern Uganda Social Action Fund (NUSAF) project to illustrate the main points in each note. We selected NUSAF for this guide because it encapsulates many facets of a standard youth livelihood program and because its impact evaluation had to grapple with many challenges. Admittedly, NUSAF is relatively large compared with many other youth livelihood projects. But as we will see, impact evaluations are also possible for smaller programs. We hope that readers will find aspects of the case study sufficiently close to their own situation.

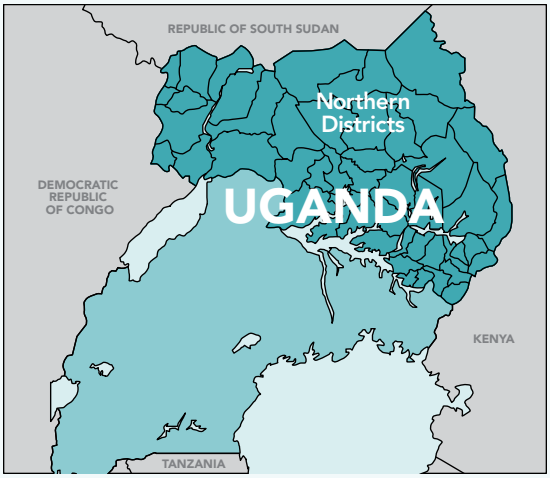
NUSAF Case Study: Background

General Information

Name of the project:	NUSAF Youth Opportunities Program
Target group:	Poor youth aged 15–35, in a postconflict region of northern Uganda
Number of beneficiaries:	8,000+
Budget:	US$1.6 million

(continued)

NUSAF Case Study: Background (cont'd)



Project Context

For two decades, most of Uganda experienced economic growth, physical security, and political stability, along with rising levels of education and health. The northern districts, however, lagged behind the rest of the country on all counts. Commercial activity has historically been located in southwestern and central Uganda due to patterns of pre-colonial and colonial development, proximity to trading partners, and availability of infrastructure.

Moreover, two decades of civil war and insecurity in the north (and in neighboring nations) destabilized the region's economy and society. Nearly all areas in the north have experienced some form of physical insecurity—armed insurgency, internal displacement, cattle rustling, and so forth. In particular, a civil war in the ethnically Acholi districts, which displaced the entire rural population of nearly two million people, has only recently concluded. As the humanitarian emergency waned, humanitarian aid phased out and national and international development assistance increased dramatically.

The Government of Uganda's Peace, Recovery, and Development Plan aspired to consolidate state authority, rebuild communities, promote peace and reconciliation, and revitalize the economy through a package of several programs. NUSAF was one of those programs.

Project Activities

The Youth Opportunities Program component of NUSAF targeted youth aged 15–35 who lived in conditions of poverty and were unemployed or underemployed. Small groups of youth self-organized, identified a vocational skill of interest and a vocational training institute, and applied to the NUSAF district technical offices for funding.

The Youth Opportunities Program had two main components.

Component 1 provided a cash transfer of up to $7,000 to local youth groups. The youth groups would use these funds to enroll in the vocational training institute, purchase training materials, and pay start-up costs for practicing the trade after graduation.

Component 2 built capacity of NGOs, community-based organizations, and vocational training institutes to respond to the needs of youth. (The length and intensity of the conflict left much of the infrastructure destroyed in northern Uganda, especially teaching institutions. By investing in these institutions, future capacity could be increased.)

Source: Blattman, Fiala, and Martinez (2011).

Overview of the Guide and How to Use It

The guide is presented as a series of short notes grouped in two major parts. The first part is about understanding the reasons for and preparing for an evaluation. The second part is about setting up an impact evaluation. Although it is important to be familiar with all parts of the process, it is not necessary to read the guide from beginning to end. Instead, each note is conceived as a self-standing chapter that can be read independently of the others, according to each reader's needs. For readers who would like to learn more about planning M&E in general, we recommend starting with part 1. Readers already familiar with M&E who would like to learn more about impact evaluation will find part 2 most relevant. The following reader's guide indicates which notes are most relevant to different types of readers.

Reader's Guide

PART I: Setting the Basis for an Evaluation The four notes in this section describe how to prepare for an evaluation.						
Note	**Description**	**Policy-makers**	**Program Managers**	**M&E Officers**	**Research and Policy Staff**	**Impact Evaluation Experts**
1	Discusses why evaluation is important and how it supports programming and organizational goals.	✓	✓			
2	Reviews some crucial questions about program design that should be answered before moving to monitoring and evaluation.		✓			
3	Presents the main steps in developing a monitoring system, which is a necessary foundation for any evaluation.		✓	✓		
4	Asks which type of evaluation best suits an individual program. The answer depends on learning objectives, the context and characteristics of the project, and available resources.	✓	✓	✓	✓	✓
PART II: Enhancing Program Learning through Impact Evaluation The notes in this section introduce impact evaluation and provide concrete guidance on its implementation in the context of youth livelihood programming.						
Note	**Description**	**Policy-makers**	**Program Managers**	**M&E Officers**	**Research and Policy Staff**	**Impact Evaluation Experts**
5	Presents the main features of an impact evaluation and explains why some commonly used evaluation methods do not fulfill the same quality criteria.	✓	✓	✓	✓	
6	Reviews tools and methods for conducting an impact evaluation and explains how they work and what they require. Also provides a decision tree to help readers reflect on which method may be best suited for their own situations.		✓	✓	✓	
7	Moves from the conceptual to the practical level, describing the major steps involved in carrying out an impact evaluation and providing practical resources. These steps cover the entire process, from initial preparations to the dissemination of results.		✓	✓	✓	✓
8	Presents tools to increase the relevance of impact evaluations. Includes an overview of the variety of impact evaluation questions, the use of mixed methods, as well as cost-effectiveness and cost-benefit analyses.				✓	✓

PART I

SETTING THE BASIS FOR AN EVALUATION

NOTE 1: Why Evaluate?

Success depends on knowing what works.

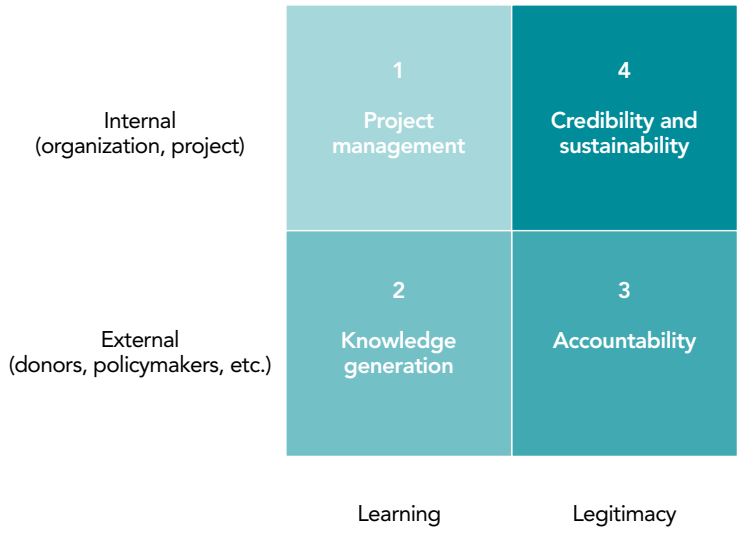
— Bill Gates

The objective of this note is to provide an overview of how individual organizations and the field as a whole benefit from the knowledge acquired from formal evaluation, particularly through impact evaluation. We argue that there are two major purposes of evaluation: learning and establishing legitimacy. For each purpose, there are internal and external audiences (see figure 1.1). Together, they yield four good reasons to conduct evaluations:

- To manage projects
- To generate knowledge
- To ensure accountability
- To strengthen our organization's credibility and sustainability

These are discussed below.

FIGURE 1.1 Benefits of evaluation



Project Management

Youth-focused interventions are inherently complex. Because we are dealing with a dynamic target group in transition biologically, socially, and legally, the interventions we put in place are highly diverse in nature and have outcomes across a range of sectors. Properly evaluating these interventions, albeit challenging, is a crucial ingredient in the recipe for success.

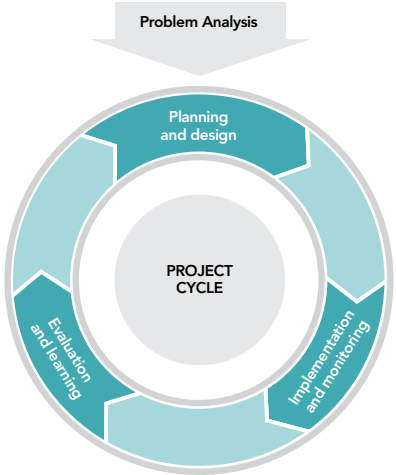
Evaluations allow us to see the true value of our work. Most of us want to know what difference our programs are making in the lives of the young people we serve. Did our project achieve the desired results? Who benefitted more, who less? Evaluations help to answer these and other questions by assessing the relevance, effectiveness, efficiency, impact, and sustainability of an intervention.

Evaluations foster learning. By assessing the design, implementation, or results of an intervention, evaluations enhance organizational learning. They allow us to identify which parts of our intervention were successful and which ones may not be working as intended. For example, an evaluation may reveal that the amount of training provided during an intervention was insufficient, resulting in low learning, or, on the contrary, was too intensive, overwhelming the students and leading to dropout. Similarly, an evaluation may help us understand unintended consequences of our project, such as an increase in parents' alcohol consumption associated with providing girls with income opportunities.

Evaluations support new planning. Evaluations provide program managers with the information we need to make strategic decisions about necessary changes in project design, planning, or implementation. Although evaluations in general (and impact evaluations in particular) produce information periodically rather than continuously, they are nevertheless valuable parts of the project cycle. Even retrospective evaluations are essentially forward looking with regard to the next generation of programming (UNICEF 1991). As illustrated in figure 1.2, evaluation applies the lessons from ongoing or terminated interventions to the planning and design of current and future programs. A well-designed evaluation helps practitioners make the necessary funding cuts to those youth programs that are not achieving their objectives, while sustaining programs that are, or could be, achieving good results. Without data from a good evaluation, the risk of reaching wrong conclusions about whether programs should continue and how resources should be allocated becomes much more significant (World Bank 2009).

> **"Having good data on why youth dropped out of training enabled us to justify providing stipends. This allowed us to bring the dropout rate from 35 percent to 9 percent."**
>
> *— Program Manager, Caribbean NGO*

FIGURE 1.2 The project cycle



Knowledge Generation

The youth livelihood field is characterized by a severe lack of sound evidence. Even if our institutions do well with regular data collection for monitoring and standard performance assessments (such as by conducting simple before-and-after comparisons or focus groups), we often fail to build generalizable knowledge that would benefit the entire field (Savedoff, Levine, and Birdsall 2006). Acquiring this knowledge typically demands impact studies that use specific methodologies to provide reliable estimates of the success of a specific intervention.

Despite the billions of dollars spent implementing youth livelihood programs, relatively few impact studies exist. For example, in a global review of the evidence of youth employment interventions, Betcherman and colleagues (2007) found only three quality evaluations of youth entrepreneurship programs. Similarly, little is known about other livelihood promotion strategies, such as second chance education, public works programming, or financial education and services for young people (see box 1.1). Even though there have been increasing efforts to build sound evidence in recent years, much more knowledge is needed.

BOX 1.1 Existing evidence on youth employment

Betcherman et al. (2007) conducted a global review of youth employment interventions and found that "only one in ten programs have evaluations which measure both net impact and cost." The types of interventions with the most severe knowledge gaps were found to be subsidized employment schemes, youth entrepreneurship, employment services, and regulatory reforms. On a regional level, evidence was particularly scarce in Asia, the Middle East and North Africa, and sub-Saharan Africa. An updated database of youth employment interventions and evaluations is available on the Youth Employment Inventory Web site (http://www.youth-employment-inventory.org).

Card, Kluve, and Weber (2009) conducted a meta-analysis of active labor market programs in the OECD. Comparing program types, subsidized public sector employment programs were found to have the least favorable impact estimates. Job search assistance programs had relatively favorable short-run impacts, whereas classroom and on-the-job training programs tended to show better outcomes in the medium-run than in the short run. The authors found that programs for youths in the OECD were less likely to yield positive impacts than untargeted programs.

Ibarrarán and Rosas Shady (2009) summarize the findings from rigorous evaluations of job-training programs in Latin America. In contrast to the evidence for developed countries, the results suggest positive effects on employment and the quality of jobs for the trainees, especially among women and the younger participants. The review acknowledges that there is still a major knowledge gap on long-term impacts of such interventions in Latin America.

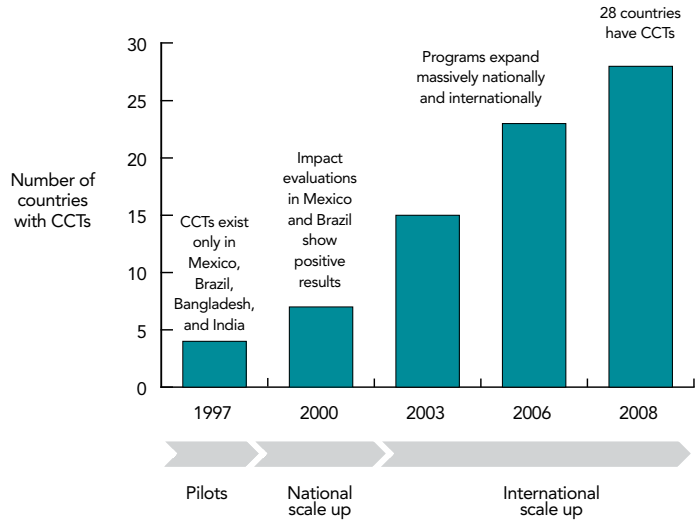
Cunningham, Sanchez-Puerta, and Wuermli (2010) summarize the state of evidence on active labor market programs for youth, classified by the constraint they are trying to address. Calling for rigorous learning and evaluation across all types of programs, knowledge gaps were found to be particularly severe for interventions such as second chance education, behavioral skills training, entrepreneurial training, public works and public service programs, technology-based job search assistance, skills certification, and microfinance, among others.

Note: See the resources section at the end of this guide for a list of completed and ongoing impact evaluations in the youth livelihood field.

The dearth of rigorous studies—despite huge demand—severely limits large-scale investments in the sector. The lack of evidence is a constraint to winning public support for youth livelihood interventions. Government officials typically want impact and cost-benefit estimates before investing in large programs. As a result of the lack of such evidence in the youth livelihood field, it is often difficult to make a convincing case in comparison to other interventions, such as infrastructure development, where much more evidence is available. Improving the evidence base would therefore also facilitate scale up and replication.

This potential has become obvious in other policy areas. The growing evidence on conditional cash transfer (CCT) programs, for instance, has enabled the international community to promote large-scale interventions in this area across the globe. Mexico and Brazil were two of only four countries worldwide with CCTs in 1997, but the evidence from their impact evaluations has resulted in a massive expansion of the model to twenty-eight countries in 2008 (see figure 1.3).

FIGURE 1.3 From evidence to policy: Conditional cash transfer programs



Sources: Fiszbein and Schady (2009); Rawlings and Rubio (2005).

The publicly funded job-training initiatives in Latin America, the Jóvenes programs, combined technical training with soft skills training, internships, and other services. Impact evaluations demonstrated measurable effects, such as an increase in employment rates and wages that reached more than 10 percent for younger and female cohorts in some countries. As a result, the Jóvenes programs were quickly replicated across the continent. Since the newer programs were also evaluated, the case illustrated the diversity of impacts that could occur in different countries and settings, highlighting that a critical mass of evaluations is always needed to be able to generalize results. Although the Jóvenes experience is still an exception in the youth livelihood sector with respect to the systematic evaluation of impacts, it gives a flavor for the possibilities for expanding the field if only we could distill better knowledge and evidence from the hundreds of interventions we are implementing every year.

Systematically building evidence about what works in strengthening young people's economic opportunities would make it possible to improve the effectiveness and efficiency of our work by bringing vital knowledge into the service of practitioners and policymakers and ultimately strengthen the entire field.

Accountability

In addition to enhancing internal and sector-wide learning, evaluation strengthens the legitimacy of our operations. Funding agencies and society are increasingly demanding accountability from development programs, and evaluations—impact evaluations in particular—can provide the needed evidence on whether a particular program achieved its desired results.

Our resources should not be taken for granted. In almost all instances, our projects are financed with public or private funds, such as official development assistance

(essentially taxpayer money) or private donations. In both cases, someone entrusted us, directly or indirectly, to use this money in the best possible way to help young people achieve a better life. The fact that we are entrusted to develop and implement a youth livelihood project means that this money is not going to be used to build rural roads, enhance an HIV/AIDS prevention project, or buy school materials. Given the scarcity of resources, it seems only natural to use evaluations to provide an honest account of our work: how the money has been used, the activities that were financed, and the results we have achieved.

We have a responsibility to ensure the best possible use of funds. Development interventions are inherently complex, and it would be illusive to expect a 100 percent success rate. On the other hand, a project does not automatically increase people's well-being simply because is well intended. In order to make sure that a specific program is doing more good than harm and that the benefits of the investment exceed the benefits under alternative uses of the resources, practitioners should always make it a priority to carefully assess the effectiveness of that intervention (Jones et al. 2009).

Credibility and Sustainability

Evaluations help increase the legitimacy of the project and the reputation of the implementing organization. This argument is not often mentioned in the literature, but in practice it may be among the most compelling reasons to conduct an impact evaluation, as it directly benefits the program and implementing organization.

Impact evaluations can enhance the credibility and reputation of our organization. Because quality evaluations are rare, they receive special attention. As a result, the simple fact that an organization or project agrees to carry out an impact evaluation already indicates good standards in programming. If the evaluation shows good results, then the payoff for the organization and program can be immense. Imagine that among the hundreds of players in the field, *you* are the one who is able to demonstrate that your method is working, that your program is successfully providing young men and women with income opportunities clearly superior to those that would have been available to them had they not participated in your program. The difference is that now you are not only able to *claim* that your intervention is effective, you are able to *prove* it. This makes a big difference in the eyes of donors and policymakers, who, prior to the evaluation, were unable to differentiate the impacts of your intervention from the alleged impacts of numerous other programs.

The ability to stand out can provide a series of benefits for both the project and the organization. Positive evaluation results can be used in advocacy and fundraising efforts to obtain greater support from donors, governments, and the general public. With greater public and political support, our project and organization can quickly become a reference in the field. This, in turn, often leads to an increase in the demand for services, and we may be expected to expand our services nationally and across borders.

Take, for example, the case of Colectivo Integral de Desarrollo in Peru. In 2003, the organization was among the very first to provide rigorous evidence that their model to promote young low-income entrepreneurs was increasing business size, improving business survival, and boosting incomes. As a result of that evidence, they received multiple awards and had no more difficulties securing funding for their programming. In fact, Colectivo is now supported by a grant program of the Inter-American Development Bank and is expanding its model to Central America and the Caribbean (see box 1.2).

International donors are increasingly looking at rigorous impact evaluations to measure the success of the programs they fund. In 2011, both the UK's Department for International Development (DFID) and the U.S. Agency for International Development (USAID) have strengthened their focus on results. See:

DFID's Business Plan 2011–2015, Section 2 *"Make British aid more effective by improving transparency and value for money"*
http://www.dfid.gov.uk/Documents/DFID-business-plan.pdf

USAID's Evaluation Policy 2011
http://www.usaid.gov/evaluation/USAIDEvaluationPolicy.pdf

"Because we could prove how youth employment improved, the government invited us to co-design an employability program under the new president."

— Program Director, Chilean NGO

A program in the Middle East provided an innovative approach to training young women for jobs as executive assistants to women entrepreneurs. The program leveraged substantial connections to the business sector and in particular to women entrepreneurs, who were interested in supporting and empowering disadvantaged young women. The initial evaluation, however, found fairly high levels of dropouts for the young women trainees once they had been placed in jobs. Further investigation as to the cause of these dropouts found that women entrepreneurs had very high expectations of these young women, but offered insufficient mentorship to them to support their success in demanding work environments. Through open dialogue with the local implementing partner, the program reframed life skills modules to better prepare youth for the demands of these jobs, and also reoriented business owners to ensure they were providing sufficient mentorship for new employees. The implementing agency was able to achieve better outcomes as a result of these mid-course corrections in the program strategy, shared these lessons with other donors, and in turn, secured additional funding to expand the program.

BOX 1.2 Benefits of conducting an impact evaluation

Response by the president of the Peruvian NGO Colectivo Integral de Desarrollo to the question "How do you think your organization has benefited from conducting an impact evaluation?"

"It improved the quality of our intervention."

"It improved the program's credibility."

"It improved the value of our brand in the eyes of donors."

"It increased demand for our services."

"We earned national and international recognition."

"We are a model institution for the replication in other contexts and countries."

Source: Dino Linares, Colectivo Integral de Desarrollo president, personal communication (January 28, 2011).

Negative Evaluation Results Are Not Necessarily Bad

We may be afraid that negative evaluation results will lead to funding cuts from our donors. Yet, evaluations that fail to confirm positive results of an intervention can be put to good use.

Negative results are unavoidable in innovative programming. Innovation and creativity are crucial to helping young people master their transition to work. Such innovation will by definition involve failures. As in any other field such as medicine, chemistry, or physics, building successful products and services requires testing, prototyping, refining, and adapting to local circumstances. Failures are a necessary step toward state-of-the-art programming.

Negative results can help improve operations. If, early on, we are able to understand the problems that may reduce the effectiveness of our intervention, then we are in good shape to build successful projects in the long run. Bad news from negative evaluation results points us toward ways of improving our programming.

Addressing negative results proactively fosters credibility. No donor or policymaker expects or believes that every project will be a great success. Disseminating findings, whether favorable or not, signals our ability to be self-critical and our commitment to continuous learning and evidence-based programming. Granted, the pressure to show results and to justify budgets can create strong incentives to report positive findings above the negative ones. But in the long term, an honest discussion of what worked and what did not is likely to yield the biggest payoff.

Key Points

1. Evaluations are first and foremost about learning for the benefit of our own project and organization. Evaluations allow us to show the true value of our work and inform the design and planning of other interventions.
2. Evaluations create a much-needed evidence base for the youth livelihood field. More and better knowledge about what works and what doesn't will help practitioners design successful interventions and convince policymakers to provide public support.

3. Evaluations provide legitimacy by holding ourselves accountable to donors and the public. Evaluations ensure the good use of taxpayer money and donations.
4. Evaluations enhance our credibility and reputation. In a sector in which robust evidence is scarce, conducting evaluations can have significant payoffs in terms of boosting demand for our services, strengthening our organization's brand, and ultimately securing sustainable financial support.
5. Evaluations don't have to show good results to be useful. On the contrary, failures foster learning. Proactively addressing and disseminating negative evaluation results will likely enhance our credibility and reputation.

The Power of Measuring Results

- *If you do not measure results, you cannot tell success from failure.*
- *If you cannot see success, you cannot learn from it.*
- *If you cannot see failure, you cannot correct it.*
- *If you can demonstrate results, you can win public support and funding.*

Source: Adapted from Osborn and Gaebler (1992).

NUSAF Case Study: Why Evaluate?

The NUSAF impact evaluation was initiated by the Government of Uganda with support from the World Bank. The primary reason for the impact evaluation was to improve program management. Seeking to estimate the impact of the Youth Opportunities Program on the livelihoods and wellbeing of youth in Northern Uganda, the evaluation was intended to inform future rounds of programming and potential scale up.

The impact evaluation was also intended to fill an important gap in understanding the effectiveness of employment and entrepreneurial skills training programs, particularly in the African context. By providing grants to obtain skills training and start-up capital for establishing productive enterprises, the Youth Opportunities Program is a hybrid of two of the most common types of employment programs. Since little is known about the effectiveness of such an approach, the evaluation would generate knowledge that could inform the entire youth livelihoods field.

Source: Blattman, Fiala, and Martinez (2011).

Key Reading

Savedoff , W., Levine, R., and Birdsall, N. 2006. *When Will We Ever Learn? Improving Lives through Impact Evaluation.* Washington, DC: Center for Global Development. http://www.cgdev.org/content/publications/detail/7973.

Notes

NOTE 2: Reviewing the Project Design

The only man who behaves sensibly is my tailor;
he takes my measurements anew every time he sees me,
while all the rest go on with their old measurements
and expect me to fit them.

— George Bernard Shaw

Imagine yourself in the following situations:

The mayor of Tripoli in Lebanon has asked your organization for technical assistance to address youth unemployment in the city. What should be done?

The manager of your country office in Rwanda is interested in self-employment programs for youth. How do you recommend proceeding?

Your strategic plan for the next three years will put stronger emphasis on young people's transition to work. Which youth employment scheme should you invest in?

Program managers are required to make difficult decisions about these and other programming issues. In order to find appropriate solutions, we need to understand the specific context and design a sound program for it. Hence, before crafting a monitoring and evaluation system, we need to make sure that our intervention itself is carefully planned: Do we have good knowledge about the needs of the people we are trying to support? Do we understand why certain conditions such as youth unemployment and social exclusion exist? Do we have a clear objective? And are we building on existing experience and evidence when designing our intervention to reach this objective?

Problem Analysis: Do We Understand the Target Group and the Local Context?

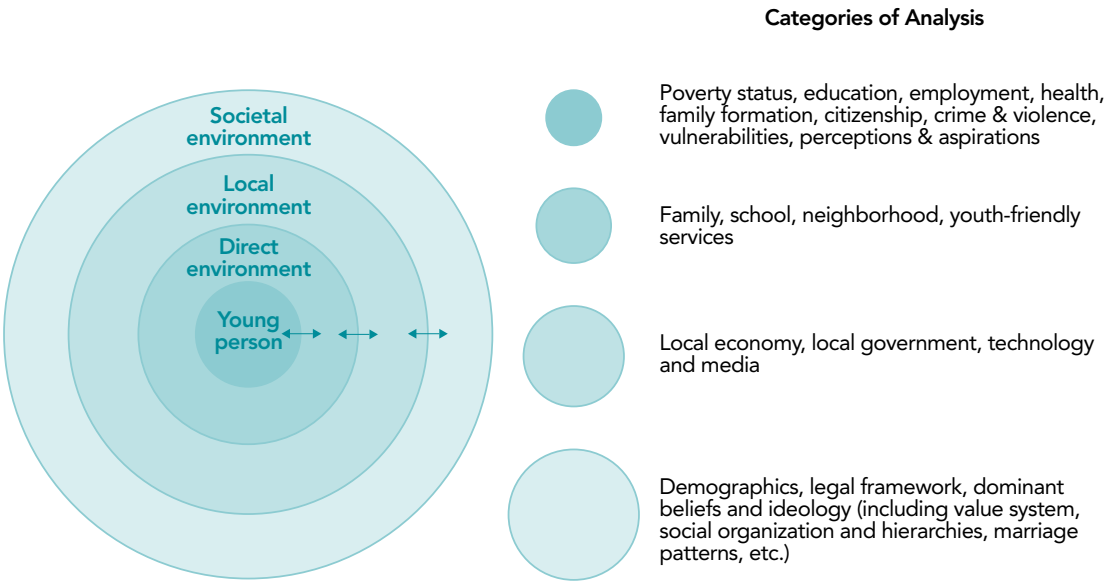
When we decide to carry out an intervention to support young people, we imply that there is a discrepancy between the status quo and what could be or should be. This gap between the existing condition and the desired condition is what we commonly refer to as a *need*. We must understand these needs before we start our intervention.

Using Cross-Sector Youth Assessments to Understand Our Target Group

Economic opportunities can rarely be understood in isolation, but are very much connected to other aspects of life. For example, employment status and income may determine one's ability to get married or form a family. Therefore, understanding the problems and needs of young people will almost always require in-depth assessments of young people's living conditions, including their socioeconomic status, behaviors and attitudes, and goals and aspirations. Youth assessments should also capture other important factors relevant to young people's transition to adulthood, such as health, family formation, and citizenship (for more information on the transitions to adulthood, see World Bank 2006).

Moreover, since young people are influenced by a wide range of factors around them, including family, peers, community, local and national institutions, and social norms, good youth assessments should also analyze the direct, local, and societal environments young people live in (see Bronfenbrenner 1979). A holistic assessment will provide a rich picture of the needs and challenges youth are facing and will therefore allow us to better adapt our intervention to local realities (see figure 2.1). Sample indicators for youth assessments are listed in appendix 1.

FIGURE 2.1 Youth environments



Analyzing young people's personal and social environments through systematically conducted interviews, focus groups, and observation will help us identify the major problem we would like to address, such as underemployment or unemployment, or the lack of access to financial services. We may also realize that limited economic opportunities are only one among many issues young people in a specific location are facing, which may suggest ways to build or adapt our intervention so that it can address more than one issue.

Equally important, the assessment will help us specify our target group. Are we interested in all youth, or only those who are out-of-school? What age range do we want to focus on? Are there gender or ethnic considerations we would like to prioritize? What are geographic areas we will target? Given our resource constraints, we are rarely able to serve every young person. A cross-sector youth assessment can help us prioritize, for example by identifying groups that are particularly vulnerable, such as school dropouts, young women, or street youth.

In conducting a rapid community appraisal of the socioeconomic profile of target youth in Jordan, the International Youth Foundation found that the level of young people's participation in civic activities in twelve target communities was extremely low (less than 4 percent). Furthermore, survey data revealed that there were very few institutions offering volunteer opportunities. When youth did participate in community service projects, they did so primarily through their schools. Focus group discussions also showed that although a "culture of volunteerism" had not taken hold in these communities, youth expressed enthusiasm for and a willingness to volunteer should opportunities be provided.

These findings helped inform the design of specific service-learning projects for out-of-school, unemployed youth. It also justified the award of grants to youth to undertake small community initiatives, which made civic engagement options more visible and accessible and ensured that they had appeal to youth. In addition, the International Youth Foundation provided training to staff of youth-serving organizations on effective development and management of community engagement and volunteer programs, which helped them to better engage youth within their communities.

Using Market Assessments to Understand the Local Economy

Given our focus on building or strengthening livelihoods, a prime component of the context analysis is the assessment of the local economy. Analyzing the local economy typically includes assessing the local labor market and assessing the market of goods and services.

Assessing the Local Labor Market

Labor market assessments seek to understand employment patterns and trends in the local economy. Common factors to analyze during such an assessment include the following (Asian Development Bank 2007, pp. 162–166):

- **Labor Demand.** Overall economic conditions; size of the formal and informal sectors; dynamic sectors or industries and geographic areas that have a demand for labor; industry trends and projections; expected number of jobs to be created; skill requirements by occupation; wage levels and earnings; working conditions; hiring practices; employer perceptions; barriers to employment based on gender, age, ethnicity, social status, religion, or other reasons; and so on.
- **Labor Supply.** Size and structure of working age population; employment, underemployment, and unemployment by gender, age, education level, urban/rural areas, sector of the economy, occupation, formal/informal, and public/private sectors.
- **Institutional and Policy Environment.** Existing labor market programs, policies, laws, and institutions, including, for example, minimum wage regulations, employment protection laws, unionization, unemployment benefits, and the like. Other aspects of interest include sectoral economic priorities defined at the national, regional, and local levels.

Assessing the Market of Goods and Services

Assessing the market for goods and services helps determine the potential for small producers to engage in sustainable economic activities and the possible distribution of roles (for example, for youth or women) in these markets (Penrose-Buckley 2007). Common market features to be analyzed include:

- **Market demands and value chains.** Existing and future gaps in terms of

consumer products and services; demand for commodities, processed products, and semifinished goods by retailers, wholesalers, or processing companies; identification of local, regional, and export markets; identification of existing market players; and other factors.

- **Market stability.** Market vulnerabilities to shocks, seasonality, and changing trends; potential restrictions to market access and the movement of people and products due to conflict and insecurity.
- **Market prices.** Price volatility of end product and supplies; potential impact of additional producers on prices; inflation; transaction costs.

Market assessments are usually carried out through a combination of analyzing existing data and surveying employers or small business holders. Interviewees can provide important insights about employment prospects in particular sectors, how hiring decisions are made, the main constraints formal and informal businesses are facing, their perceptions of young people, and more. This information, in turn, can inform the diagnostic and program design (see following sections). See box 2.1 for examples of youth and market assessments.

The International Rescue Committee's (IRC) LEGACY Initiative in Liberia focuses on bolstering community-driven education programs and expanding market-driven vocational training opportunities to young women and traditionally excluded youth. In order to ensure the curricula were market-driven, IRC conducted two assessments.

A labor market survey identified marketable trades that have potential employment opportunities for youth. The assessment tool was a ten-page questionnaire that combined multiple choice and open-ended questions to garner better understanding of employers' existing and potential recruitment needs.

In addition, IRC conducted a rapid market survey to help two vocational training centers identify products and services for potential school-based businesses. The assessment included questionnaires with retailers, customers, and suppliers in local markets.

Among other things, the assessments helped IRC understand the needs of young girls compared with those of boys, which helped increase the number of girls that would eventually enroll in the training centers.

Source: Beauvy-Sany et al. (2009).

BOX 2.1 Sample youth and market assessments

In-depth youth and market assessment:

Peeters et al. 2009. *Youth Employment in Sierra Leone*. Washington, DC: The World Bank. http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2009/02/27/000334955_20090227091204/Rendered/PDF/476090PUB0Sier101Official0Use0Only1.pdf

Rapid appraisal youth assessments:

International Youth Foundation. 2010. *Building on Hope: Findings from a Rapid Community Appraisal in Jordan*. Baltimore: IYF. http://www.iyfnet.org/document/1059

International Youth Foundation. 2011. *YouthMap Senegal—Youth Assessment: The Road Ahead*. Baltimore: IYF. http://www.iyfnet.org/document/1820

Diagnosis: What Are the Determinants Influencing Youth Outcomes?

What follows the youth and market assessments? Let's assume we found that the young people in the country, region, or city we work in are disproportionately affected by unemployment and underemployment. Most youth ages 16–24 are neither in school nor working. Girls seem to be particularly affected. So, what should we do? What intervention can we propose?

In fact, these questions are premature. Before we think about an intervention, we need to know *why* these young people are unable to find work or start their own business. What prevents them from entering the labor market and making a living for themselves and their families? What constraints are they facing?

Imagine we put in place technical training courses targeting young women. The training could provide artisan skills and computer literacy based on a model our

organization has successfully implemented elsewhere. But what if technical skills weren't the problem to begin with? What if the real issue for these young women was a lack of knowledge about how to look for and apply for a job, combined with social constraints that discourage young women from working outside the household? If we have too little knowledge or the wrong assumptions about why young people are unable to find work, chances are our project will not address the root problem and therefore will not be successful. In such a case, monitoring and evaluation will only confirm the obvious.

During project design for a youth employment and enterprise development program in Indonesia, a local NGO did not fully diagnose the underlying psychosocial problems and economic constraints that youth were facing as a result of a destructive past earthquake in the area. The earthquake not only took numerous lives and displaced more than 50,000 people but also destroyed livelihood facilities. Vital counseling services were not available to youth after their traumatic experiences; in fact, the need for ongoing psychosocial support was not even recognized. Although monitoring visits revealed that some job creation was successful, the program failed to meet its overall targets as youth continued to suffer from depression and struggled to adopt new technologies or take other steps that could have made their small enterprises more profitable.

Understanding Long- and Short-Term Barriers

It is crucial to understand the underlying constraints that may limit young people's access to the labor market and to income-generating activities. Here, we conceptualize these constraints as those that probably cannot be addressed within the timeframes of most programs (long-term), and those that can (short-term).

Long-term constraints. Institutional and macroeconomic issues take time to address and are unlikely to be influenced by individual local projects, whose time frames are typically three to five years. Yet, long-term constraints are important to consider because they represent the larger context of our intervention. Box 2.2 presents the MILES framework, an overview of structural determinants to job creation.

BOX 2.2 The MILES framework

Macroeconomic and political stability. Entrepreneurs require a sound macroeconomic framework in which to expand their business and create new jobs.

Investment climate, institutions, and infrastructure. Firms will expand and create formal sector jobs when the costs of doing business (from regulation, heavy tax burden, and poor infrastructure) are low and predictable.

Labor market regulation and institutions. Sound regulations are crucial for both the employer and the worker to engage in a productive, long-term working relationship.

Education and skills. High productivity jobs are invariably based on good formal education and require appropriate skills for all age groups.

Social protection. A strong and balanced social protection scheme protects the income of workers from shocks to employment.

Source: World Bank (2007c, pp. 8–10).

Short-term constraints. Given that it is difficult to change most structural barriers to employment and livelihood creation, it is usually more realistic for development practitioners to focus on the constraints that can be addressed in a shorter period of time. Among those, five major categories stand out (Cunningham, Sanchez-Puerto, and Wuermli 2010):

- **Supply-side constraints:** Youth lack job skills relevant to the local market, including basic literacy and numeracy skills, technical skills, behavioral skills, or entrepreneurial skills. They may also face non-skills related constraints, including psychosocial issues, which may affect their employability (Rossiasco et al. 2010).
- **Demand-side constraints:** Employers express low demand for youth labor because of macrolevel effects, such as slow job growth, as well as microlevel effects, such as employer discrimination.

- **Business creation constraints:** Constraints to youth entrepreneurship include lack of access to financial capital, land, or social networks.
- **Labor market intermediation constraints:** Young people often lack relevant and accurate information about job openings and about qualifications in demand, or they cannot adequately communicate their skills to potential employers.
- **Social constraints:** Social norms or customs may limit skills development or labor market entry for particular groups, such as girls, indigenous youth, and others.

Table 2.1 provides an overview of possible constraints. In practice, the challenge is to determine which ones are the most relevant in our local context and to prioritize them accordingly. Each subpopulation of interest will likely face a different set of constraints. For example, young women in rural Rwanda live in a low-growth economy, will lack skills, face severe employer discrimination, and be limited by gender norms, while low-income men in urban Chile may be most constrained by information about job opportunities, difficulty in communicating competencies to potential employers, and by a mismatch of technical or soft skills (Cunningham, Sanchez-Puerto, and Wuermli 2010). The short list of constraints for our specific target population needs to be identified through youth and market analyses, as described above.

TABLE 2.1 Overview of short-term constraints for young people in the labor market

Constraint		Description	Information Sources
Supply side: Job-relevant skills and other supply-side barriers	Insufficient basic skills	• Literacy and numeracy are the foundation of communication and further skills development processes. • Young people in postconflict settings may be particularly affected.	• Skills assessments, including skills certification (e.g., PISA, TIMSS) • Education system assessments (SABER) • Statistics about educational achievement • School curricula (what is being taught?) • Existing employer surveys • Market/sector assessments
	Insufficient or mismatch of technical skills	• Trade- or job-specific skills range from manual skills to computer literacy.	
	Insufficient behavioral skills	• Behavioral skills—or soft skills—consist of a range of qualities such as motivation, problem solving, communication, time management, and the ability to work with others. • Behavioral skills are increasingly valued by employers around the world.	
	Insufficient entrepreneurial skills	• The creativity to invent or adopt a new product or process and the business skills to market the idea are essential for both employees and the self-employed.	
	Barriers not related to skills	• Other constraints that may affect a young person's ability to accept or look for work, such as transportation cost, child care responsibilities, etc. • A young person's health, especially mental health, can influence their employability. For example, depression or anxiety can affect their behavioral skills.	• Youth assessment • Specialized health and mental health assessments or studies

TABLE 2.1 (CONT'D) Overview of short-term constraints for young people in the labor market

Constraint		Description	Information Sources
Demand side: Lack of labor demand	Slow job growth	• Too few new jobs are created (small formal sector). • Slow growth often results from economywide factors such as a difficult investment environment or from external shocks, such as natural disasters, war, or a sudden change in the global economy.	• GDP data (by sector) • Labor market statistics (including share of formal/informal economy) • Existing employer surveys • Market assessment
	Employer discrimination	• Employers may have prejudices against youth, believing, for example, that young people are less reliable, less trustworthy, or less skilled, than older people. • In addition to prejudices, hiring preferences may be made along gender, racial, ethnic, or religious lines.	• Existing employer surveys • Market assessment • Sociological studies
Business creation: Firm start-up constraints	Lack of access to financial, natural, and social capital	• Limitations to self-employment may include a lack of entrepreneurial skills as well as inadequate access to money, land, or business networks.	• Market assessment • Banking/microfinance statistics (loan products, collateral requirements, etc.) • Property rights
Intermediation: Job-search constraints	Job matching	• Youth often lack the established networks to find out about available jobs.	• Youth assessment • Market assessment
	Signaling competencies	• Youth may have the right skills, but it may be difficult to communicate these skills to potential employers (e.g., through prior experience or certificates).	• Youth assessment (regarding diplomas, certificates, etc.) • Market assessment
Social constraints	Excluded group constraints (ethnicity, gender, etc.)	• Local customs and social norms may deter certain groups of people from participating in the labor market. • Occupational segregation may occur along racial, ethnic, or religious lines.	• Market assessment and/or existing employer surveys (hiring preferences) • Human rights reports • Anthropological studies

Source: Adapted from Cunningham, Sanchez-Puerto, and Wuermli (2010).

Objectives and Design: What Do We Want to Achieve and How?

In light of the specified problems, target group, and specific barriers to better economic opportunities, we can formulate program objectives and select among possible interventions. Clearly defining what we want to achieve will help us think about the end results of our program, communicate with donors and stakeholders, manage the intervention, and monitor and evaluate our work.

Setting the Project Development Objective

The first step is to define our project development objective.[1] The project development objective represents what we want to accomplish, the intended or planned result of our intervention. Several tasks can help us develop our objective, such as clearly specifying the target group, the magnitude of the expected changes, and the time period (see figure 2.2). The more concrete the objective, the easier it will be to track progress against it.

FIGURE 2.2 How to develop project objectives

Based on the specific problems and constraints identified:

1. Specify the expected **result**—what we expect to achieve, not what we want to do.
2. Specify the **target population**.
3. Specify the **direction and magnitude** of the expected changes.
4. Specify the **time period** in which expected changes will occur.
5. Make sure the objective is **measurable**.
6. Make sure the objective can be **attributable** to the intervention.

A common mistake when defining our project development objective is to focus on what we will do, instead of what we intend to achieve (see point 1 in figure 2.2). If the ultimate reason for our intervention is to improve the living conditions of young people, then that should be reflected in our project objective. The way we achieve this goal—for example by providing psychosocial support, training, seed capital, or other services—is the "how to" and not the actual objective. Box 2.3 assesses three examples of a project development objective.

1 Organizations use various terms to label their project development objectives, such as project goal, final goal, or purpose.

BOX 2.3 Defining a project development objective

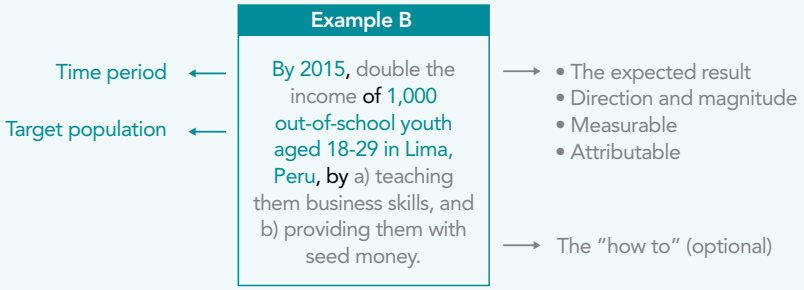
Example A	Example B	Example C
1,000 Peruvian youth trained in business skills.	By 2015, double the income of 1,000 out-of-school youth aged 18-29 in Lima, Peru, by a) teaching them business skills, and b) providing them with seed money.	By 2015, reduce youth unemployment in Lima, Peru, by 10%.

Assessment

This sample objective lacks several necessary details (time period, exact target group, etc.) which make it too generic. Moreover, it does not refer to an expected result, but rather to a service that will be delivered. The fact that youth will be trained does not necessarily translate into an improvement of their situation, such as an increase in knowledge, employment status, or income. The project objective should go beyond that.

This example fulfills all the requirements for a good project objective. It is concrete and refers to a measurable improvement in the target group's living conditions.

This objective, while specific and measurable, is impossible to attribute to our project. A single intervention of limited scope will not be able to bring about the desired high-level change, as youth unemployment will be influenced by a variety of factors.



Setting Institutional Objectives

In addition to defining the project development objective, we may also be interested in defining institutional objectives. Institutional objectives are linked to our intervention, but they may not directly refer to our primary target group. For example, institutional objectives can be internal to our organization, such as learning lessons from the project in terms of design and implementation. Institutional objectives may also refer to the project environment, such as building partnerships, fostering political will, or improving stakeholder involvement. All of these are important and should be defined from the outset of the project.

Defining the Intervention

With a clear goal in mind, we can define the scope of the intervention that will lead us to achieving our stated objectives. Naturally, the choice of the program should directly result from the specific barriers identified in the previous section; that is, we should choose an intervention that explicitly addresses the underlying causes that hinder young people's abilities and opportunities to make a decent living for themselves and their families.

Evidence-Based Programming

A crucial element in developing an intervention is reviewing the existing knowledge about various program alternatives. For example, to address business start-up constraints for young people, we may want to implement a program to promote youth enterprises. But what exactly should the intervention look like? Assume we were able to confirm that financial constraints are the major obstacle to starting a business. Should the program provide grants or loans? Should it target younger or older youth? The less or the better educated? And will financial support be enough, or should it be combined with other support services, such as training, mentoring, and business development support?

To answer these difficult questions, program managers will certainly benefit from looking at the existing evidence base. Many times, we (or the organizations we work for) tend to favor certain types of projects based on our predispositions and prior experience. Yet, in order to develop high-quality projects, it is important to consider the existing theoretical and empirical knowledge about youth livelihood programming. (The resources section at the end of this manual includes references to academic journals, databases, and past and ongoing impact evaluations). If the available evidence confirms our inclination, then we can make a strong case for a specific design. If, instead, existing knowledge points to serious limitations of an intervention, then it will save time and money to incorporate the lessons learned into the new initiative.

Table 2.2 provides examples of interventions that have a good track record based on previous impact evaluations or positive monitoring data. Building on those programs will help design better and more credible interventions. A thin or missing evidence base does not mean that a proposed intervention is doomed to failure. In fact, innovative approaches will by definition lack a track record. However, when we carry out interventions that lack a good evidence base, we should always be aware of their probationary nature and not take positive results for granted. This is where rigorous evaluation will be especially important.

[Tip]

The Youth Employment Inventory (www.youth-employment-inventory.org) is a one-stop source for ongoing and past youth employment interventions. The dynamic database allows browsing and filtering by type of intervention and evaluation, enabling users to search for available evidence on a specific type of project.

TABLE 2.2 The menu of evidence-based interventions, by constraint

Constraint	Intervention with Strong Evidence	Intervention with Mixed Evidence
Insufficient basic skills	• Information about the value of education	• Second chance education programs
Technical skills mismatch	• Training "plus"/comprehensive programs Information on returns to technical specialties	• On-the-job training
Behavioral skills mismatch	n/a	• Behavioral/life skills training
Insufficient entrepreneurial skills	n/a	• Entrepreneurial training
Slow job-growth economy	• Wage or training subsidies	• Public service programs • Labor-intensive public works
Employer discrimination	• Affirmative action programs	• Subsidies to employers who hire target groups • Employee mentoring
Lack of access to financial, natural, or social capital	• Comprehensive entrepreneurship programs	• Microfinance
Job matching	• Employment services	• Technology-based information sharing
Signaling competencies	n/a	• Skills certification • Training center accreditation
Excluded group constraints (ethnicity, gender, etc.)	• Target excluded group's participation in programs • Nontraditional skills training • Safe training/employment spaces for specific groups	• Adjusted program content/design to account for excluded group specific needs

Source: Adapted from Cunningham, Sanchez-Puerta, and Wuermli (2010).

[Tip]

Taking a holistic view of youth development, livelihood promotion strategies should be understood in a broader context of what young people need to successfully transition to adulthood. For example, the Search Institute's Developmental Assets framework presents forty internal and external assets of young people that can be strengthened to foster positive youth development.

For more information, see: http://www.search-institute.org/developmental-assets

Given the economic, social, institutional, and administrative diversity within and across countries and the specific needs of the target group, all the interventions in table 2.2 will not necessarily be feasible in a specific context. Assess whether the preconditions exist in the target country or labor market, and, if they don't, whether the program design can be adjusted to make the intervention feasible (Cunningham, Sanchez-Puerta, and Wuermli 2010).

The Link Between Program Design and Evaluation

Finally, it is important to recognize that there are important linkages between program design and evaluation. As we have seen in note 1, one of the major roles of evaluation is to support learning and, in turn, planning. The usefulness and feasibility of the evaluation therefore very much depends on the quality of the original program design. Keep the following points in mind:

- **Evaluation does not make up for poor design.** Later evaluation does not replace early thinking. A well thought out program design based on existing research and experience is the best we can do for a successful program.
- **The evaluation strategy will depend on the knowledge gaps identified during the design stage.** Knowing the evidence base and identifying potential knowledge gaps are important factors in choosing the right evaluation strategy later on. For example, impact evaluations will be particularly valuable for innovative and untested programs that provide an opportunity to fill in global knowledge gaps.
- **The right program design can facilitate evaluation.** Some programs are easier to

evaluate than others. For example, if an impact assessment is not planned during the design stage of the program, the tools available to conduct the evaluation may be severely constrained (see note 6). In turn, choosing clear, fair, and transparent targeting criteria, such as random assignment for oversubscribed programs or eligibility scores, can significantly ease the evaluation. Thus, if there are multiple acceptable ways of delivering a particular program, it may be wise to plan ahead and choose the design that also suits the evaluation.

Key Points

1. Since the usefulness of monitoring and evaluation ultimately depends on the quality of the original project design, we must ensure high standards in the planning and design of our interventions.
2. To design quality projects, we must understand youth and the context they live in. This requires cross-sectoral youth and market assessments that capture the complexity of environmental factors that influence young people's wellbeing and opportunities.
3. It is crucial to diagnose the underlying factors that impede young people's access to employment and income. Without knowing what exactly limits their opportunities, it is impossible to design an intervention that addresses the relevant constraints.
4. When designing an intervention to achieve the stated project development objective, consult existing theoretical and empirical evidence to increase the likelihood of success and prevent costly mistakes.

NUSAF Case Study: Reviewing the Project Design

Problem Analysis

High levels of youth unemployment and underemployment are persistent problems that appear at the top of the policy agenda for many governments in low- and middle-income countries. This is true also for the Government of Uganda, which is looking for ways to mitigate the chances of future conflict arising in the north of the country.

Diagnosis

In Africa in general, and in northern Uganda in particular, there are almost no formal sector employment options for people due to a lack of private businesses. Given the lack of employment opportunities combined with low levels of skills and barriers to starting a business, the Youth Opportunities Program decided to focus on a comprehensive entrepreneurship program that would provide vocational skills training, cash grants, and other support services.

(continued)

NUSAF Case Study: Reviewing the Project Design (cont'd)

Objectives and Design

The Youth Opportunities Program had the following main objective. By 2010, it sought to increase employment for at least 8,000 youth aged 15–35 in Northern Uganda by promoting skills-based enterprises and building the capacity of training facilities (the desired magnitude of the employment effect was not specified).

In addition to the main objective, the program targeted a number of secondary objectives, such as improving the young people's social interactions in their communities and decreasing the psychological distress caused by the recent conflict. More broadly, the program aimed at contributing to the overall wellbeing of youth and their households, improving health and quality of life, providing sustainable economic growth, and, as a result of these, reducing the likelihood of future conflicts arising in northern Uganda.

Source: Blattman, Fiala, and Martinez (2011).

Key Reading

Bidwell, K., Galbraith, C., et al. 2008. *Market Assessment Toolkit for Vocational Training Providers and Youth*. New York: Women's Commission for Refugee Women and Children and Columbia University School of International and Public Affairs. http://www.womensrefugeecommission.org/docs/ug_ysl_toolkit.pdf

Cunningham, W., Sanchez-Puerta, M. L., and Wuermli, A. 2010. "Active Labor Market Policies for Youth: A Framework to Guide Youth Employment Interventions." Washington, DC: The World Bank. http://siteresources.worldbank.org/INTLM/214578-1103128720951/22795057/EPPNoteNo16_Eng.pdf

UN Capital Development Fund. 2011. *Listening to Youth: Market Research to Design Financial and Non-Financial Services for Youth in Sub-Saharan Africa*. New York: UNCDF. http://www.uncdf.org/english/microfinance/uploads/other/Listening%20to%20Youth-YouthStart%20Market%20Research.pdf

Women's Refugee Commission. 2009. *Building Livelihoods: A Field Manual for Practitioners in Humanitarian Settings*. New York: WRC. http://www.womensrefugeecommission.org/docs/livelihoods_manual.pdf

Notes

NOTE 3: Establishing a Monitoring System

What gets measured gets done.

— Tom Peters

A good evaluation is impossible without a good monitoring system. Moreover, designing a good monitoring system will likely enhance the overall quality of our project design and facilitate project management. This note summarizes the key steps for building a monitoring system that should be followed in any project, whether or not an evaluation will take place. As we will see, at minimum, each project should have the following monitoring tools in place:

- A results chain
- A logical framework
- A process to collect and analyze information and apply findings

Why Do We Need a Monitoring System?

Monitoring provides internal and external information on a continuous basis to inform program managers about planned and actual developments. When irregularities or inefficiencies are detected, they can be corrected in a timely manner. Monitoring involves collecting and analyzing data to verify that resources are used as intended, that activities are implemented according to plan, that the expected products and services are delivered, and that intended beneficiaries are reached (Savedoff, Levine, and Birdsall 2006). Effective monitoring needs to be part of any project, regardless whether the project will be evaluated.

Monitoring also provides the foundation for evaluating an intervention. In fact, a good evaluation is hard to conduct without proper information about actual implementation. If no reliable information about the progress and quality of implementation is available, then any evaluation will run the risk of misinterpreting the reasons for success or failure of the project.

The challenges in monitoring an intervention are to

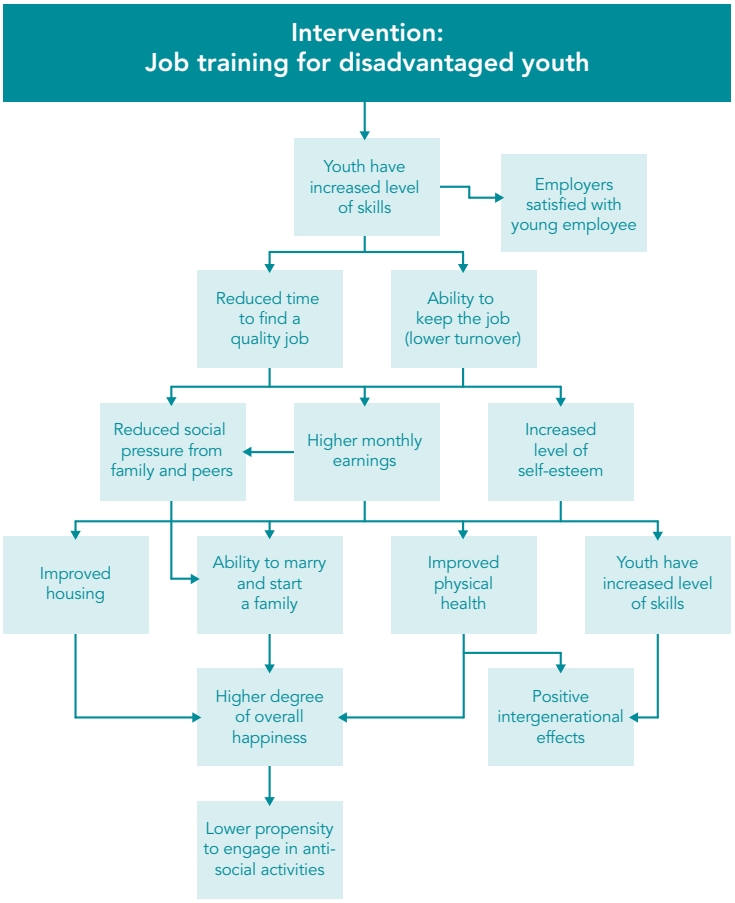
- define the *logic of the intervention*, which includes setting goals beyond the project development objective on all levels of implementation and results.
- identify *key indicators, data collection mechanisms*, and *assumptions* that can be used to monitor progress against these goals.
- establish a *monitoring and reporting system* to track progress toward achieving established targets and to inform program managers and other stakeholders.

Defining the Logic of the Intervention

The Link Between Project Design and Project Theory

Encapsulated in any program design is a theory of change. As discussed in note 2, usually there is an expectation that a project will help improve the living conditions of our target group by addressing a specific set of barriers and constraints these young people face. That is, we have a set of assumptions about how and why particular project activities will foster positive change. Why do we believe that training youth will result in better labor market outcomes? Why do we believe that supporting youth enterprises will reduce poverty? To confirm the relevance of our intervention, the theory behind it has to be clear (see figure 3.1).

FIGURE 3.1 Basic intervention theory of a youth livelihood project



Practitioners should articulate a theory of change for every intervention. Ideally, it is developed at the beginning of the project design phase, when all relevant stakeholders can be brought together to agree on a common vision for the project, its concrete objectives, and the steps necessary to achieving those objectives (Gertler et al. 2011). According to Taylor-Powell (2005), using a theory of change helps both the project manager and the evaluator by

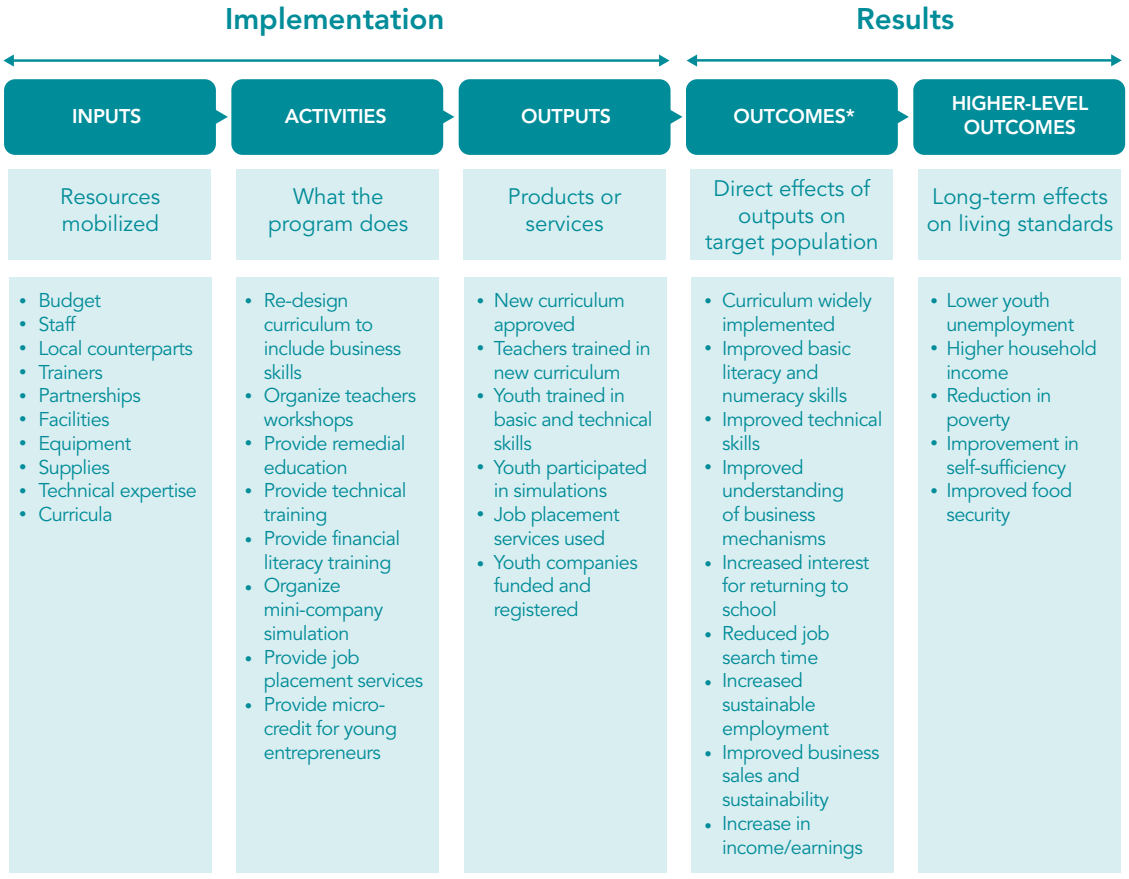
- increasing understanding about the program and providing a common language.
- helping to differentiate "what we do" from "what we want to achieve."
- improving planning and management.
- identifying important variables to measure.
- providing a foundation for in-depth evaluations.

Turning the Theory Into a Results Chain

In practice, a theory of change can be applied in a variety of way. Common applications include logic models, logical frameworks, outcome models, or results chains. The idea is always the same: to provide stakeholders with "a logical, plausible outline" of how the planned intervention can lead to the desired results (Gertler et al. 2011, p. 24; see figure 3.2). As a result, they present a sequence of events that connects the elements under direct responsibility of the project (resources used, activities implemented, and

products and services provided) with the expected outcomes and higher-level objectives of the program.

FIGURE 3.2 Components of a results chain and examples



** Level of Project Development Objective*

[Definition]

A **results chain** is a sequence of resources, activities, and services provided that are expected to influence the direct and long-term effects on our target population.

Our planned implementation process includes the following categories a program manager is directly responsible for:

- **Inputs**—the resources available to the project, including budget, staff, partners, and equipment.
- **Activities**—the actions, processes, techniques, tools, events, and technologies of the program. Describe these activities with an action verb (*provide, facilitate, deliver, organize,* etc.).
- **Outputs**—the products and services provided that are directly under the control of the implementing organization. They indicate if a program was delivered as intended. Outputs are typically expressed as completed actions (*trained, participated, used, funded,* etc.).

Our intended results describe all of the program's desired effects under the following categories:

- **Outcomes**—the short- to medium-term effects (usually within several months and up to two years) on the beneficiary population resulting from the project outputs. These may include changes in attitudes, knowledge, and skills, which can often be relatively immediate effects, as well as changes in behaviors, status, and the like, which may take more time. The key outcomes targeted should be those defined in the project development objective. Outcomes are typically expressed at an individual level and indicate an observable change (*increased, improved, reduced*, etc.).
- **Higher-level outcomes**—the long-term project goals usually relating to overall living standards. They can be influenced by a variety of factors and are typically not under the full control of the program. This level of the results chain is also often labeled "impacts." We prefer the phrase "higher-level outcomes" to avoid confusion about the specific meaning of "impact" in the context of impact evaluation (see note 5).

[Tip]

Though not absolutely necessary, it is often a good idea to also include your institutional objectives and underlying activities in the results chain.

Constructing a Results Chain

Define the Level of Observation

Both in terms of the implementation and results, we may want to look at more than individual youths. In fact, we may also be interested in outputs or outcomes at the household level, the group or facility level (schools, vocational training centers), or the village/community level.

Consider the Diversity of Possible Outcomes

Youth livelihood interventions can affect a multitude of outcomes, including, but far beyond, outcomes that directly relate to economic opportunities and the labor market. Depending on the intervention, it may be useful to target and measure a range of outcomes if these fit the project logic and objectives. Common outcome categories include the following:

- **Psychosocial development**—measures of a young person's mind, emotions, and maturity level. Outcomes can relate to self-esteem, identity, trust, isolation, or psychological wellbeing.
- **Skills**—levels of basic knowledge in literacy and numeracy; technical competencies in a specific trade (artisan, mechanics, accounting, customer services); life skills (communication, teamwork, critical thinking, self management); and entrepreneurial skills (creativity, business skills).
- **Employment and labor market**—beneficiaries' use of time (between school, wage employment, self-employment, unemployment, casual labor); job characteristics (type of employer or business, number of hours worked, earnings); and business characteristics (profits, number of employees, business survival, loan repayment rate).
- **Use of financial services**—beneficiaries' access to financial services and behaviors related to banking, saving money, debt management, budgeting, and overall financial well-being.

- **Risky behaviors**—attitudes and behaviors relating to alcohol, tobacco and drug use, reproductive health (e.g., unprotected sex, HIV/AIDS), crime and violence.
- **Family formation**—attitudes and behaviors concerning age of marriage and the desired and actual number of children.
- **Citizenship**—young people's preferences and actions with respect to voting in local or national elections, engaging in the community (such as through club membership or volunteering), and assuming leadership roles.
- **Investments in human capital**—changes in educational status (has returned to school or would like to return to school), amount of money spent on education or health (for herself or others), and intergenerational contributions (e.g., immunization and growth monitoring for own children).
- **Other**—additional outcomes may relate to consumption and nutrition patterns, asset creation, mobility and migration, as well as household and community relations.

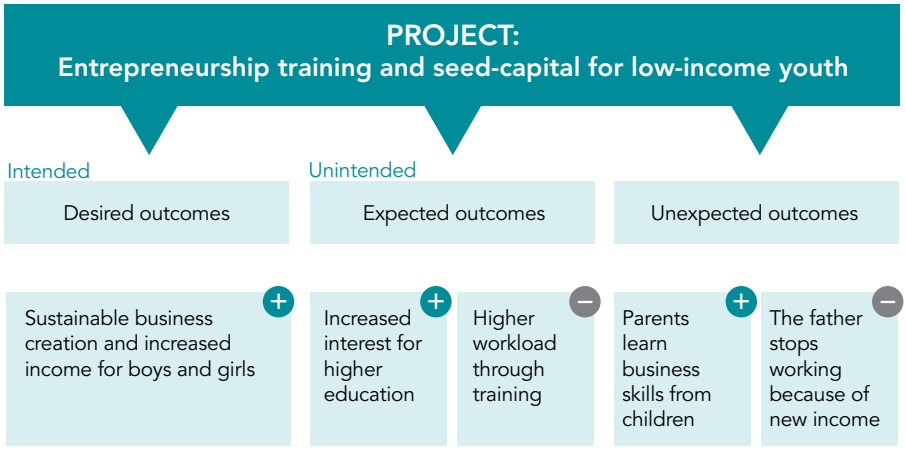
In the early 2000s, the Population Council and Save the Children implemented the Ishraq Program in rural Upper Egypt, establishing girl-friendly spaces to impart life skills, build social networks, and foster leadership and self-confidence. As it turned out, program benefits went beyond the targeted out-of-school adolescent girls and extended to the parents of participants. Girls conveyed information from their classes to their mothers, including health information. Additionally, observing their daughters' participation in public life had a strong impact on mothers' perceptions of their own place in the public sphere. Thanks to their daughters' involvement in Ishraq, mothers realized that they, too, had a right to access public services.

Source: Brady, Salem, and Zibani (2007).

Take Unintended Outcomes Into Account

Our project objective reflects the major desired outcome of the intervention. Yet, development projects are complex and our intervention may have *unintended* effects. Some of these unintended effects may be expected, while others are unexpected and surprising. Both expected and unexpected outcomes may be positive or negative (see figure 3.3). It is important to include these potential outcomes (see major categories above) in the results chain and to label them accordingly in order to realistically capture the full logic of the intervention and provide the basis to track all mechanism at work.

FIGURE 3.3 Intended versus unintended outcomes



Source: Adapted from Hempel (2006).

For example, there may be spillover effects from our intervention because participants transfer knowledge to family or community members who, in turn, may also benefit indirectly. We certainly would like to capture this effect. On the other hand, there may be negative effects that are not expected: In an entrepreneurship project, for example, some youth may find themselves trapped in debt because their business did

not survive. In other cases, where youth are generating higher incomes thanks to our intervention, family members may stop working or may use the additional income to increase unhealthy behaviors such as alcohol and tobacco consumption. Again, we want to know whether these effects are actually at play. Doing research about similar projects can often help identify the range of potential positive and negative outcomes.

Avoid Redundant Activities or Outputs

When developing our results chain, we may identify activities that have little to do with our main project objective. In the interest of a well-defined and efficient project, such activities and outputs that are not crucial should be dropped.

[Tip]

If tracking unintended outcomes risks overwhelming the results framework, project teams may choose to focus monitoring on the intended outcomes and use evaluations to capture the extent of unintended outcomes.

Identifying Key Indicators, Data Collection Tools, and Assumptions

Once we have a results chain, how do we know whether what has been planned is actually happening? One of the biggest challenges in developing a monitoring system is choosing what kind of information best reflects whether we are reaching our objectives. We now try to identify appropriate indicators, data collection tools, and assumptions for each level of objectives, from inputs to higher-level outcomes. A logical framework provides a useful matrix to capture all these elements (see table 3.1).

Step 1: Identifying Indicators

Indicators answer the question "How will I know?" Indicators are

- key aspects of (or proxies for) the element that we want to measure, even though they may not necessarily be fully representative.
- tangible signs that something has been done or that something has been achieved; they are the means we select as markers of our success (Shapiro 2003).

Indicators are a crucial element of a monitoring system because they drive all subsequent data collection, analysis, and reporting. Without a clear set of indicators, monitoring or evaluation activities lose their capacity to compare a program's actual progress with what was projected and agreed upon (Gosparini et al. 2004).

TABLE 3.1 Example of a logical framework for a school-based entrepreneurship program

		Step 1	Step 2			Step 3
	Objectives	**Indicators and Targets**	**Information Source**	**Frequency**	**Responsible Party**	**Assumption**
Higher-Level outcomes	• Lower youth unemployment • Higher household income	• Local unemployment rate (%) • Household income ($)	• Employment statistics (ministry, city level) • Household survey	• Yearly	• Program team	• New skills are demanded and rewarded by labor market
Outcomes	• Curriculum widely implemented • Better understanding of business • Improved soft skills • Improved employability • Increased interest for higher education	Within six months of completing the program: • 500 schools use new curriculum • 50% more correct answers on business knowledge post-test • 70% students satisfied with new curriculum • Teacher and parent perceptions of soft skills improve by 30% • Time spent searching for a job falls 50%, and employer satisfaction increases 30% • 5% increase in university enrollment	• Interview with official education authority • School test results • Focus group with teachers and parents • Employer survey • Regional school enrollment statistics	• Bi-yearly	• Program team (interviews, data collection) • consultant (survey, focus group)	• Curriculum accepted by local school authorities • Quality of teaching • Youth can attend school regularly
Outputs	• New curriculum approved • Teachers trained • Youth trained in business skills	By the end of the program: • New curriculum approved by ministry • 500 teachers trained • 10,000 youth completed the training	• Program data	• Bi-monthly	• Program team	• Teachers willing to be trained • youth can attend training
Activities	…	…	…	…	…	…

Note: In the interest of practicality we have omitted the activities and inputs categories, which would usually be included in the logical framework.

Selecting Indicators for All Levels of the Results Chain

Even when our focus is on the results of the intervention, it is important to track implementation indicators so we can determine whether the project has reached its intended beneficiaries and whether it has been carried out as intended. Without these indicators all along the results chain, an evaluation will identify only whether the predicted outcomes were achieved, but it will not be able to make a connection between the level of success and the quality of program execution. Table 3.2 illustrates examples of such indicators along the results chain.

TABLE 3.2 Examples of indicators

Category	Sample Target	Example of Indicators
Input	Two trainers and facility within budget of US$10,000	• Two trainers skilled, equipped and deployed • Cost of program in U.S. dollars within desired budget
Activity	Provide life skills training for youth (20 hours)	• Number of training hours delivered • Number of youth participating by age, gender, level of education • Date by which training was provided
Outputs	100 youth participated in training	• Number of youth who finished the training (by age, gender, level of education)
Outcomes	Increased knowledge of effective communication	By the end of the program: • Number and percentage of youth able to express ideas clearly measured against a predetermined test score card • Number and percentage of youth with improved verbal and nonverbal communication skills measured against a predetermined test score card • Number and percentage of youth who report feeling comfortable approaching employers
Higher-Level Outcomes	Increased household income	• By 2015, average monthly household income increased by 20% compared to baseline

Defining good *outcome indicators* requires particular attention. As discussed above, the outcomes of youth livelihood interventions can be very diverse and are not limited to labor market outcomes. We therefore need to choose indicators (psychosocial development, skills, employment, etc.) among all appropriate domains. The precise domains to be measured depend of course on the goal and focus of the intervention and learning objectives to be addressed. For example, for a life-skills intervention, it may be useful to measure skills, labor market outcomes, and risky behaviors. A job placement support project, instead, may be entirely focused on labor market outcomes.

[Online Resource]

Selected outcome and output indicators

http://www.iyfnet.org/gpye-m&e-resource1

Specifying Indicators

Bring in other stakeholders. Choosing indicators without the proper involvement of primary internal and external stakeholders can lead to a lack of ownership on their part (Kusek and Rist 2004). Collaborate with local partners and stakeholders in the community to arrive at a mutually agreed set of goals, objectives, and performance indicators for the program.

Choose the right number of indicators. Since indicators are only proxies, it is common to define several indicators for each element in the results chain, especially regarding outcomes or higher-level outcomes. However, choosing too many indicators will unnecessarily complicate our monitoring system and increase the burden for data collection, analysis, and reporting. It is important to identify the one to three key indicators that best reflect each element in the results chain.

Respect quality standards. Even though there are no absolute principles about what makes a good indicator, the commonly cited SMART characteristics can be useful (Gertler et al. 2011, p. 27). SMART indicators are

- Specific, to measure the information required as closely as possible,
- Measurable, to ensure that the information can be readily obtained,
- Attributable, to ensure that each measure is linked to the project's effort,
- Realistic, to ensure that the data can be obtained in a timely fashion, with reasonable frequency, and at reasonable cost, and
- Targeted to the objective population.

Our selection of indicators will be in part determined by our ability to collect data on them. If an indicator cannot be measured or the information is not available, then it cannot serve its purpose to reflect progress of our objectives. If we are not able to collect data for an indicator we chose, we have to replace it.

Establish a baseline. The baseline tells us the value of an indicator at the beginning of, or, ideally, just prior to, the implementation period. Knowing the baseline value of our indicators allows us to define realistic targets and track future progress against the initial situation. For example, if we want to monitor participants' incomes over time, data from our program registration forms may tell us that the average monthly income of participants at the time they enter the program is $100. This is our baseline value that can serve as a comparison for how incomes will develop during and after our intervention.

Define targets. If indicators are not specified in terms of time frame, quantity, and quality, we cannot be completely sure about being on track and reaching our objectives (Cooley 1989). For example, if the desired outcome is increased household income, our indicator may be monthly earnings in U.S. dollars. Then, the target may be set at a 30 percent increase (quantity) from formal sector employment (quality) within three years (time frame). Each indicator should have no more than one target per specified period. If setting firm numerical targets is too arbitrary, then targets can also be expressed as a range.

Ensure consistency. Although it is not always possible, in order to ensure consistent monitoring over time, we should make an effort to retain the indicators that were agreed upon before the start of the project. That said, it is not uncommon to add new indicators and drop old ones as we modify the program design or streamline the monitoring system. However, it is essential to remember the original objectives of the project. Monitoring and evaluation must be truthful. If we find that our project will not achieve its original goal but will instead achieve some other goal (which may be of even greater value), we must acknowledge that in our reporting. Indicators accepted at the beginning of the intervention should not be changed unless objective criteria exist to justify the change.

Table 3.3 provides examples of indicators for youth livelihood interventions at all levels of the results chain. Sometimes it is possible to use pre-defined indicators. However, it is important to consider their relevance to the specific project. Some may need to be adapted to fit or supplemented with others that are more locally relevant.

[Tip]

It is usually a good idea to pilot indicators during the early phases of an intervention before establishing them as integral part of the monitoring system. This will highlight how well they work and whether they are capturing all the information the project manager and other stakeholders are interested in.

Outcome to be measured: *Improved employability of youth aged 18–24*

Bad indicator: Youth will find jobs more easily than they could before the intervention

Good indicator: Number and percentage of youth aged 18–24 who have at least two job offers that pay above minimum wage in their field of training within three months of completing the program (compared to zero job offers before)

TABLE 3.3 Examples of indicators for youth livelihood projects

Type of Project	Input	Activities	Outputs	Outcomes	Higher-Level Outcomes
Training and skills development	• Budget allocation and expenditure (in U.S. dollars) • Amount and share of matching funds raised • Number of program staff by level • Number of local facilitators under contract • Number of local organizations who provide in-kind contributions	• Number of workshops offered • Number of training hours • Number of youth screened/enrolled • Number of employers offering internships • Number of internships available	• Number and percentage of youth who attend at least 80% of the training • Number of certificates awarded • Number of youth placed in internships • Average length of internships completed (in weeks)	• Number and percentage of youth who are satisfied with the program • Number and percentage of youth reporting an improved ability to think critically and solve problems • Number and percentage of youth receiving follow-up jobs offers after internship • Percentage of local employers providing job opportunities for young people	• Household income (in U.S. dollars) • Local youth unemployment rate (%) • Levels of individual/household food consumption (including fruit and vegetables) • Number and percentage of youth who report that their house/apartment has basic infrastructure (running water, electricity, etc.) • Number and percentage of youth who report reduced levels of conflict in the previous year
Subsidized employment (e.g., public works and public service programs)	Same as above	• Number of workfare projects by type and location • Number of municipalities providing public work/services	• Number of beneficiaries employed in each activity • Number of temporary jobs created (by type and sector)	• Number and percentage of youth who transitioned to formal employment within X months • Days and hours worked per week (by type of activity) • Average hourly/daily/monthly wage	Same as above

TABLE 3.3 (CONT'D) Examples of indicators for youth livelihood projects

Type of Project	Input	Activities	Outputs	Outcomes	Higher-Level Outcomes
Employment services (e.g., job placement support)	Same as above	• Number of career counseling services created (in labor offices, in schools, etc.) • Number of job counseling session offered • Number of career and job fairs organized	• Number of youth participating in job placement services • Number and percentage of youth matched with employers • Number of companies and youth participating in local career/job fair	• Number of job interviews per beneficiary • Number and percentage of youth who are employed X months after the intervention • Number and percentage of youth who retain employment for at least X months	Same as above
Youth enterprise and entrepreneurship	Same as above	• Number of business plan competitions organized • Number of hours of support services provided • Average number of hours of mentoring provided per week/month	• Number of youth submitting complete business plan • Number of youth enterprises supported annually • Number and percentage of youth talking to their mentor at least once every two weeks	• Number and percentage of youth who started a new business • Number and percentage of businesses registered • Total sales last week/month • Number of jobs created • Percentage of profits reinvested	Same as above
Youth-inclusive financial services	Same as above	• Number of workshops organized for participating financial institutions • Micro-loan scheme for young entrepreneurs launched • Youth-targeted savings account created	• Number of staff trained in partner financial institutions • Number of business loans issued to young people (by type of enterprise) • Average loan size • Number of youth saving accounts opened	• Annual repayment rate • Amount of current savings (1) in bank account, (2) with savings group, (3) in all other locations • Number and percentage of youth who put aside savings as soon as money comes in • Number and percentage of youth who report greater satisfaction with financial situation	Same as above

Step 2: Data Collection

The selection of indicators to be used for our monitoring system depends not only on the project structure and objectives, but also on the availability of data and on the time and skills requested for their collection. Data refers to information of all types, not just quantifiable information.

In Pakistan, the Mennonite Economic Development Associates monitors its rural economic development projects with an SMS reporting system. Women microentrepreneurs and small enterprise owners submit daily or weekly sales reports via their personal mobile phone.

Select a Data Collection Method

There are two broad methods of data collection: quantitative and qualitative.

Quantitative methods aim to provide an objectively measurable picture of a situation in some strictly predetermined ways. They provide information about the population of interest in closed-form and quantitative dimensions, including demographic, socioeconomic, or other characteristics. They are usually based on standardized structured instruments that facilitate aggregation and comparative analysis. Common examples include tests, surveys, and censuses. Conducting quantitative methods requires skills in statistics.

Qualitative methods aim to provide an understanding of how and why people think and behave the way they do. Qualitative methods seek to understand events from stakeholder perspectives, to analyze how different groups of people interpret their experiences and construct reality. Common examples of qualitative methods include unstructured or semi-structured interviews, focus groups, and direct observation of participants. Conducting qualitative methods requires training in anthropology or sociology, as well as training in the administration of specific evaluation tools. Qualitative methods tend to be quicker to implement than quantitative methods, and are often less expensive.

The rules governing statistics are transparent and comparatively easy to follow, requiring little independent judgment from the analyst. As a result, quantitative methods usually achieve higher standards of reliability and validity compared with qualitative methods. In contrast, the interpretation of qualitative data is a matter of judgment. As a result, qualitative methods are more difficult to generalize. Given the advantages and limitations of both categories, a mixture of qualitative and quantitative methods (mixed-methods approach) is often recommended to gain a comprehensive view of the program's implementation and effectiveness. Table 3.4 provides an overview of common data collection techniques.

With the rapid development and expansion of information and communication technologies, there is a increasing array of technology-based solutions that can be used to facilitate data collection. This includes the use of mobile phones and other mobile devices to implement surveys, Web-based tools, mapping instruments, and other multimedia solutions.

[Online Resource]

Overview of ICT-based data collection tools

http://www.iyfnet.org/gpye-m&e-resource2

TABLE 3.4 Overview of data collection methods

Method	Description	Use	Advantages	Limitations
Administrative and Management Records	Documents that provide information on project management processes	To examine the effectiveness of project management or strategy implementation	• Provides information on process that is difficult to obtain through other means	• Program specific, not generalizable • Dependent on reliable management records systems
Field Visits (combination of observation and interviews)	In-depth examination of a specific site or location	To monitor and understand context	• High level of detail • Access to observational data	• Program specific, not generalizable • Highly dependent on access to appropriate field sites
Key Informant Interviews	In-depth data collection method with highly informed individuals	To obtain specific and highly detailed information on a specific issue or set of issues	• High level of detail • Can address unanticipated topics • Has flexibility to explore issues in depth • Can capture a range of stakeholder perspectives	• Program specific, not generalizable • Quality is highly variable based on interviewer skills and interviewee comfort
Focus Groups	In-depth data collection method with informed members of a specific subpopulation (e.g., women, youth, elderly, workers)	To obtain specific and highly detailed information on stakeholder perspectives on a specific issue or set of issues	• Same as for key informant interviews • Allows for interaction with and among participants	• Program specific, not generalizable • Quality highly dependent on group dynamic (e.g., participants can be influenced by moderator or dominant group members) • Interpretation challenges • Time-consuming analysis
Direct Observation	Method to collect data through direct observation (e.g., classroom observation), information is recorded in a log or diary	To obtain naturalistic data	• High level of detail from a neutral observer • Provides information on actual behavior rather than self-reported behavior	• Not generalizable • High potential for observer bias • Interpretation and coding challenges

Method	Description	Use	Advantages	Limitations
Review of Official Records	Official documents that provide background information or historical data on certain phenomena	To examine underlying processes or historical trends/data for certain phenomena	• Provides information that may be difficult to obtain through other means • Inexpensive	• Possible access restrictions • Must verify validity and reliability of data • Data may not be exactly what is needed
Mini-Surveys	Brief questionnaire/survey that collects limited data set	To obtain quantitative data on a limited number of people or issues	• Faster and less expansive than household surveys	• Limited scope and therefore usually not representative
Household Surveys	An extensive set of survey questions whose answers can be coded consistently	To obtain information on a large number of respondents regarding their socioeconomic status, demographic data, consumption patterns, etc.	• Provides in-depth information on population of interest • More generalizable than mini-surveys • May be designed to collect data of specific interest	• Expensive • Requires special expertise to ensure validity • Difficult to persuade people to respond to long questionnaire
Panel Surveys	A longitudinal study in which variables are measured on the same units over time	Same as for household surveys, with particular interest in measuring changes over time	• Same as for household surveys • Can capture dynamics over a period of time	• Same as for household surveys • May have problems with participant retention over time
Census	Survey for an entire population	To obtain a complete data set on a specific population	• Generalizable • Typically available from official sources	• Expensive • Time consuming • Infrequent or dated

Sources: Adapted from Baker (2000); Creswell (2008).

[Tip]

Use quantitative methods when

- numerical or generalizable data are required to convince decision makers.
- you need statistically representative information about the target population, their situation, behaviors, and attitudes.

Use qualitative methods when

- "how and why" questions need to be understood; that is, when quantitative data need to be explained by motivation and attitudes affecting behaviors.
- participatory approaches are favored.

Data collection mechanisms are more or less suited for different levels of the results chain. Input and process indicators will rely primarily on management and project records that illustrate the use of resources and the implementation of activities. Direct observation and field visits can provide data for output indicators, for instance, the number of small businesses created. Measuring outcomes often requires a combination of formal surveys that provide reliable quantitative information as well as qualitative methods such as key informant interviews or focus groups to understand the underlying mechanisms of whether and how certain effects were achieved. Finally, since higher-level outcomes usually relate to broader changes outside the full control of the project, official statistics can be useful when they are available for small geographic areas (such as municipalities) and can be disaggregated by sociodemographic characteristics.

Define the Frequency and Timing of Data Collection

The interval of monitoring activities will depend on the monitoring purposes. As a rule, the higher the level of the results chain, the less frequent we will need to collect data, but the more difficult it usually becomes to obtain accurate information.

To illustrate the optimal frequency of data collection, let's imagine a job-training program that lasts for three months. To run the training effectively and efficiently, we need information about how many resources we are using (in terms of budget, staff time, materials, etc.) and how our activities are implemented (the number of hours of training offered every week, the number of participants, and so on). This information about our inputs and activities may need to be collected fairly frequently, let's say every two weeks.

[Tip]

The timing of data collection should be planned against local realities so that collection does not impose a burden on an individual or a family. Data should not be collected when youth are taking school exams, for example, or when young people's labor is needed during particular agricultural seasons.

Assessing our output (the number and the composition of beneficiaries that are actually being trained) would probably be done periodically, say, every month, although this information will rely on attendance data that may have been collected on a daily level.

Whether the training had any effect on outcomes (youth's knowledge and ability to find employment) will only become clear after the training is over. Short-term effects, such as an increase in knowledge, may be measured at the end of the training, while effects that take longer to manifest—such as whether jobs were secured—would be measured three to six months after the intervention.

Finally, higher-level outcomes such as increases in household income and positive spillover effects are usually unlikely to materialize in less than a year (depending on the local labor market) and would therefore be measured only in long intervals.

Define Who is Responsible for Collecting the Data

It is important to clearly define data collection responsibilities. Failing to define responsibilities will likely result in failing to collect the data. In practice, different types of monitoring will fall under the responsibility of different actors, both in the field and at headquarters. The following people are likely to collect data:

- Program managers
- Local project team members or M&E officers
- Local implementing partners (e.g., teachers, training providers, loan officers)
- Beneficiaries
- Other local stakeholders (including parents and community members)
- Volunteer enumerators (e.g., university students)

- External consultants
- Survey firms

While defining the responsibilities for collecting the data, clarify what happens to the information once collected. Integrate data collection plans with procedures for storing, aggregating, and analyzing the data to guarantee that those who need the information have timely access to it (see Monitoring and Reporting System, below).

To learn more about participatory monitoring and evaluation, consult Sabo Flores (2008), Powers and Tiffany (2006), and Gawler (2005).

Step 3: Articulating Risks and Assumptions

What are the key factors that could diminish the potential effects of our project, and what steps can be taken to mitigate them? In any project there are factors that we cannot control that will affect the success of our intervention. These could include such factors as weather, political stability, the local security situation, and support from local stakeholders. A good understanding of these factors is essential for project design, and also for M&E.

Identify Assumptions During the Design Phase

We can identify assumptions by thinking of the factors critical to reaching our objectives on each level of the results chain and what could affect these factors (see table 3.5). Sometimes, a first set of assumptions may already have been formulated in the *risk* section of our project proposal (Development Marketplace 2008).

[Tip]

Be mindful of conflicts of interest when assigning responsibilities for collecting and reporting information. For example, teachers or training providers may have an incentive to cheat with respect to recording outputs (such as the number of hours of training conducted) or outcomes (such as the number of youth who improved their test scores or found a job). To ensure data reliability, we recommend (1) using neutral observers to ensure independent monitoring, and (2) verifying the accuracy of information provided, at least sporadically, through unannounced site visits or other means.

For an example how photo monitoring improved teacher attendance and reduced the need for monitoring visits in India, see http://www.povertyactionlab.org/evaluation/encouraging-teacher-attendance-through-monitoring-cameras-rural-udaipur-india

TABLE 3.5 Examples of assumptions and project responses

Category	Potential Assumption	Under Our Control? Yes/No
Input	• Trainers willing to work in project area can be found • Employer association ready to partner	• Yes, but not hired yet • Yes, memorandum of understanding already signed
Activity	• Electricity available for training location	• No, but no problems in recent months
Output	• Youth can attend training and don't have to work to support family	• No, but vouchers given to compensate for lack of income
Outcome	• Training is relevant to labor market needs and delivered with high quality	• Yes, employer surveys carried out and trainers' performance will be monitored
Higher-Level Outcomes	• Local economy (including market prices and wages) remains stable	• No, but predictions are good

Assumptions that are not under our control should be inserted in the results matrix at the level they influence. In general terms, inputs and activities are more likely to be under the project's control than outcomes and higher-level outcomes.

Making unrealistic assumptions regarding some key elements of the program can seriously impede the success of the intervention, and should thus be avoided in any circumstance. This can happen when we overestimate the resources we have at hand, lack knowledge about beneficiaries and local context, and are unable to adequately assess external risk factors such as insecurity or opposition from local government. (Development Marketplace 2008).

Monitor Assumptions During Project Implementation

In order to provide an early warning system on potential constraints as well as on possible solutions, assumptions should be closely followed. Monitoring assumptions allows us to know how they may be affecting project implementation and results, and therefore can help us explain deviations from our objectives and take corrective measures.

Establishing a Monitoring and Reporting System

Planning

After a full logical framework with indicators, data collection tools, and assumptions has been developed, the following tasks will help you to prepare for monitoring.

Design necessary instruments. Data collected systematically with well-designed instruments will enable reports to be generated quickly and reliably. Instruments should be piloted with a germane population during development, and results from the pilot exercise should guide the design of subsequent instruments.

Develop procedures to protect young people. Although not always required by national governments, professional norms dictate that data collection activities be administered in such a way to protect the rights and interests of participants. The exact nature of these procedures is subject to local requirements, but, at a minimum, the following are encouraged:

- Create instruments and interviewer training procedures that ensure the anonymity of young research participants.
- Obtain signed informed-consent forms that include details of the project and the potential risks associated with participation. These forms also clearly explain the rights of participants, such as the right to drop out of the data collection process whenever they like. Obtain oral consent from people who cannot read.
- Obtain informed consent from the parent or guardians of people who are under the legal age of consent, people who are developmentally disabled, and other vulnerable populations. If such a person is not available to consent, avoid collecting data on the vulnerable individual.

For more detailed guidance, see the section Human Subjects Protection in note 7.

[Tip]

Make sure that the instruments used capture various types of contact information (physical address, email, telephone number) from the respondent and also from friends and family who can help locate the highly mobile youth later on. Using social media channels such as Facebook can also help to communicate with and keep track of young people.

[Online Resource]

Sample survey instruments, some of which include consent forms

http://www.iyfnet.org/gpye-m&e-resource11

Collect the data according to the chosen methods. To the extent possible, existing processes such as participant registration or attendance records should be leveraged in order to minimize the data collection burden to staff and respondents.

Develop the database. If the data collected is complex, it may be beneficial to employ an experienced database manager to develop codes and procedures that allow multiple users to query the data and derive results with a little bit of training. A variety of database systems are appropriate, and the project should select a software program that provides a balance of analytical sophistication and user-friendliness.

Aggregating and Analyzing Information

The methods for aggregating and analyzing findings are highly dependent on the methods one employs to monitor a project or intervention. Therefore, decisions on how to use monitoring data should start very early in the design process. The project team must decide upon the best ways to organize these data and conduct effective and efficient

analysis. To facilitate analysis and reporting in bigger programs, it may be advisable to set up a Management Information System that connects all databases used by different program units.

For qualitative data, it is often ideal (albeit logistically challenging) to employ computer-based qualitative analysis software. There are many brands to choose from (such as Atlas.ti, NVivo, or MaxQDA), and each work in similar ways. Software for qualitative analysis allows the user to import all relevant documents (such as transcripts from interviews and focus groups, project documents, and photographs) and then apply a set of predetermined codes. Depending on the sophistication of the user, the codes can function as an organizing tool (grouping all like topics from various sources together) or allow sophisticated analysis that examines for relationships within these topics. The team should choose the software that meets their needs in terms of staff experience and costs.

For quantitative data, when resources allow, it is often best to use a number of systems. One should be a relational database, such as Microsoft Access. Relational databases allow for an easy investigation and display of data along a number of different variables. Typically, however, the analyses performed in relational databases are fairly descriptive in nature, providing measures of central tendency (e.g., means, modes, medians, standard deviations). If the project demands, and the instruments are designed and administered in such a way as to allow for more sophisticated analysis, the monitoring staff may want to use a statistical software package such as SPSS or Stata. In addition to commonly available statistical software packages that are based on the hard drive of a single computer, there is also an increasing use of "cloud"-based data management and analysis systems, which allow a large team to collaborate on monitoring and analytical tasks.

Learning and Decision Making

Monitoring has little value if we do not learn from and act on the data that results from the analysis. Being in a constant cycle of action and reflection helps to remember that situations change, that the needs of project beneficiaries may change, and that strategies and project activities need to be reconsidered and revised. Organizations and projects stagnate when they don't learn, and rigorous monitoring forces us to keep learning (Shapiro 2003).

According to Shapiro (2003), translating learning into action entails

- looking at the potential consequences of our analysis on our program.
- listing options for action.
- discussing the options with internal and external stakeholders, reaching consensus, and obtaining a mandate to take action.
- sharing adjustments and plans with the rest of the organization and, if necessary, with our donors and beneficiaries.
- implementing the plan.
- monitoring the effects.

[Definition]

A **Management Information System** is the combination of computer technology, people, and procedures put in place to collect, organize, and analyze information in order to support decision making. It allows for centrally managing large amounts of data and for comparing indicators by beneficiary characteristics and over time.

In 2011, Youth Business International (YBI), a network of more than thirty-four independent youth entrepreneurship programs around the world, began implementation of a cloud-based global Operations Management System (OMS) for monitoring purposes. The OMS allows YBI members to track and analyze a broad range of key performance indicators relating to organizational efficiency and outcomes. The quality of a member's loan portfolio and the success of their entrepreneurs' businesses can be assessed against factors such as the sociodemographic characteristics of the entrepreneur, the mentoring and training delivered, and the terms of the loan. The platform helps increase accuracy and facilitates real-time aggregation of information by the central YBI network team.

Reporting

Typically, the higher the standing of our audience in an organization's hierarchy, the less we need to provide a lot of detail and explanation in communicating our findings. Presenting clear messages substantiated by aggregated data and concise information tends to be more appropriate for high-level audiences, who are mostly interested in the big picture. We can tailor the format of our reports to suit each audience (see table 3.6).

TABLE 3.6 Tailoring reports to audience

Target Audience	Format	Timing/Frequency
Project Staff	Oral presentation and written summary statistics at team meetings	Weekly
Management Team	Written reports and oral presentation	Monthly
Partners	Oral presentation and written summary statistics	Monthly
Donors	Depends on donor requirements. Usually short written reports highlighting project progress, issues experienced, outcomes and impact, efficacy of intervention/ strategy, etc.	Quarterly/biannually

Monitoring data should always be reported in comparison with their baseline and target values and presented in a "simple, clear, and easily understandable format" (Kusek and Rist 2004, p. 133). Visual tools, such as graphs, charts, and maps can be very useful in highlighting key data and messages.

Resources

Monitoring systems can be expensive. In addition to fixed costs (computing hardware and software, staff) there are also variable costs that include training local enumerators, contracting outside consultants, and publicizing findings (see table 3.7). It is important that a project's M&E system is properly budgeted and accounted for in any strategic plan. It is often the case that when the costs are realized, program managers hesitate to spend significant resources on an M&E system, which appear to be at the expense of intervention activities. Yet, without suitable monitoring systems, a program runs the risk of underperformance or failure, with little awareness of these problems. Also without monitoring, we may not be able to seize those opportunities where great successes are being realized. At the end of the day, monitoring systems are critical to project management and a crucial component of any intervention.

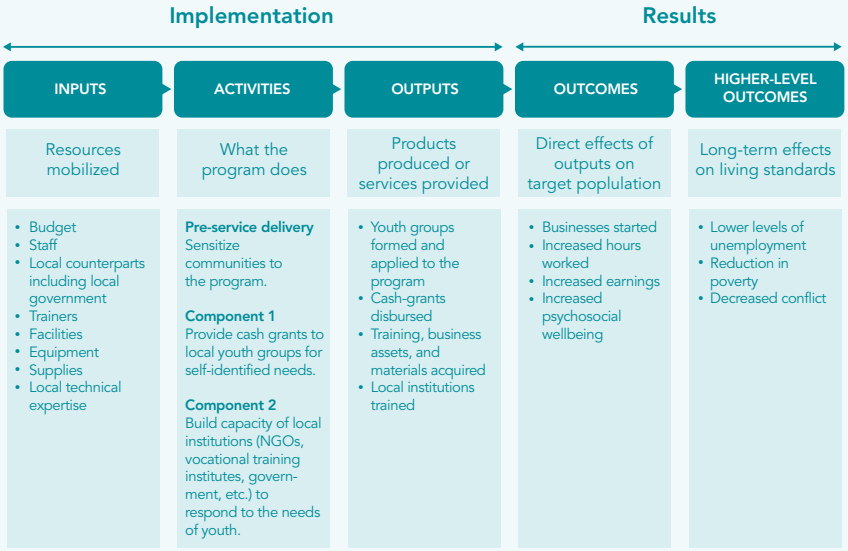
TABLE 3.7 Typical components of a monitoring budget

Fixed Costs	
Staff Cost	• Headquarters: Percentage of an M&E coordinator's time to manage M&E system. Can range from 10 percent to 100 percent, depending on project size. • Locally: Typically 50–100 percent of a local M&E officer's time to manage implementation of M&E activities, plus junior support staff.
Equipment	Computers, voice recorders, cameras, etc.
Software	Licenses for quantitative and qualitative analysis tools
Variable Costs	
Training	Capacity building for staff, enumerators, community members, etc.
Travel	Travel from HQ staff to the field for periodic check-ins and technical assistance. Local travel to field sites to ensure standardized implementation of M&E activities
Data collection and Analysis	Contracting of third party vendors such as survey firms
Consultants	Contracting of external experts for specific tasks
Printing	Instruments, reports, etc.

Key Points

1. Every intervention must have a solid monitoring system to be able to continuously track implementation and results, regardless of whether the project will be evaluated.
2. Program managers and key stakeholders need to jointly develop a results chain to clearly specify the logic of the intervention and identify key indicators, data collection mechanisms, and assumptions.
3. The monitoring system provides continuous information on the direction, pace, and magnitude of change. It also allows us to identify unanticipated developments in the project or its environment. This provides the foundation for knowing whether an intervention is moving in the intended direction and makes good monitoring critical to effective project management.
4. Monitoring data is descriptive and does not necessarily explain why and how certain changes are taking place. It also does not provide the basis for attributing the observed changes to the intervention; that is, it does not prove that changes are taking place because of our program.

NUSAF Case Study: Monitoring System



In order to build the foundation for interpreting the results of the impact evaluation, it was crucial for the NUSAF program to have good information about whether the Youth Opportunities Program was implemented as intended. NUSAF therefore used a mix of quantitative and qualitative tools to track activities and outputs. For example, since cash grants were disbursed to youth groups through the central government, youth were asked whether they actually received the funding. This information was then compared with government records.

The program also tried to understand the distribution and use of the money within the group. Because the money was intended for training, materials, and tools, NUSAF tracked attendance rates, the number and value of their tools and materials, whether they began a business, and whether they were still operating the business.

Although this information did not provide answers regarding the impact of the program, it helped program officials, monitoring staff, and the evaluators to understand whether the program was delivered as planned and how it may have affected participants. This understanding would also help during the analysis of evaluation results, for example to explain why some participants may have benefited from the program to a different extent than others.

Source: Blattman, Fiala, and Martinez (2011).

Key Reading

Donor Committee for Enterprise Development. 2010. *The DCED Standard for Measuring Achievements in Private Sector Development. Control Points and Compliance Criteria.* Version V.
http://www.enterprise-development.org/page/measuring-and-reporting-results

Kusek, J. Z., and Rist, R. C. 2004. *Ten Steps to a Results-Based Monitoring and Evaluation System: A Handbook for Development Practitioners.* Washington, DC: The World Bank. See chapters 2–6.
http://www.oecd.org/dataoecd/23/27/35281194.pdf

Taylor-Powell, E., Jones, L., and Henert, E. 2003. *Enhancing Program Performance with Logic Models*. University of Wisconsin–Extension, Program Development and Evaluation. http://www1.uwex.edu/ces/lmcourse/

W. K. Kellogg Foundation. 2004. *Logic Model Development Guide.* Battle Creek, MI: W. K. Kellogg Foundation. http://www.wkkf.org/knowledge-center/resources/2006/02/WK-Kellogg-Foundation-Logic-Model-Development-Guide.aspx

Notes

NOTE 4: Choosing the Right Type of Evaluation

The most serious mistakes are not being made
as a result of wrong answers.
The truly dangerous thing is asking the wrong question.

—Peter Drucker

Although a good monitoring system is critical to knowing whether our intervention is moving in the intended direction, it does not necessarily answer the question how or why changes are coming about, nor does it prove that any observed changes in outcomes are the result of our intervention. To complement the information we obtain from our monitoring system, we need evaluations. Evaluations are periodic assessments of the relevance, efficiency, effectiveness, impact, and sustainability of our intervention. The type of evaluation best suited for our project will depend primarily on our information needs. Therefore, the first step to any evaluation is to define what we want to learn. These learning objectives as well as our operational context, in turn, will determine which type of evaluation is right for our program.

What Is the Purpose of the Evaluation?

As a first step to deciding if an evaluation is necessary and which design should be chosen, it is crucial to clearly define what we want to get out of the evaluation. *What decision will be informed by the evaluation and what kinds of information are needed to make that decision?* Do we want to know more about how well our programs are being implemented, whether our programs are meeting their objectives, or whether our beneficiaries are actually better off as a result of our intervention? As program managers and evaluators, we must first establish our questions and learning objectives and then select the most appropriate evaluation tool to provide the necessary information (Karlan 2009).

Broadly speaking, evaluations address three types of questions (Imas and Rist 2009):

- **Descriptive questions** seek to describe processes, conditions, organizational relationships, and stakeholder views (*What is going on in our project?*).
- **Normative questions** compare what is taking place to what should be taking place. They compare the current situation with the specific objectives and targets that have been defined (*Has our project been implemented as intended? Is it performing as expected?*).
- **Cause-and-effect questions** examine outcomes and try to measure what difference an intervention makes. They ask whether objectives have been achieved as a result of our project (*What is the impact or causal effect of our program on outcomes of interest?*).

Which of the above questions we should ask is ultimately up to us, based on the specific intervention.

Organizing our questions. In practice, we may have many questions across all categories that we would like to answer. An effective way to organize all the possible evaluation questions is through our results chain (see table 4.1). In fact, if a good monitoring system is in place (see note 3), there should be consensus around our project logic in terms of implementation and results, which in turn makes it easier to identify the critical learning objectives along all stages of the intervention. Descriptive and normative questions can relate to all levels of the results chain; however, cause-and-effect questions specifically refer to outcomes and higher-level outcomes.

TABLE 4.1 Examples of evaluation questions

	Inputs	Activities	Outputs	Outcomes	Higher-Level Outcomes
Descriptive	• How does the cost of the program compare to similar interventions? • What are the qualifications of service providers? • What are other ongoing interventions?	• Do youth know about the program and how they qualify to join? • What delivery mechanisms are being used? • To what extent does the program implementation differ by site?	• How many youth participate (by age, sex, etc.)? • Who drops out? • What services are used the most?	• Are participants satisfied with the program? • Are there any observable changes in participant knowledge, attitudes, etc.? • How many program participants find employment within 3 months?	• Is local youth unemployment rising or falling? • Are household incomes evolving?
Normative	• Do we spend as much as we have budgeted? • Are the staff and financial resources adequate? • Is the program duplicating other efforts?	• Is the process for selecting participants fair and equitable? • Is the program implementation delayed? • Are operational manuals being followed?	• Do we achieve the desired gender balance in participants? • Will we reach the goal of training 5,000 youth per year?	• Does participant income increase by 20%, as planned? • Do 80% of beneficiaries find a job within 3 months of graduation, as required? • What, if any, are the unintended positive or negative effects?	• Are more households becoming self-sufficient? • Are more households reaching food security?
Cause-and-Effect	n/a	n/a	n/a	• As a result of the job training, do participants have higher paying jobs than they otherwise would have? • Does including internships increase the effectiveness of technical training offered? • Does the program affect boys and girls differently?	• Does the project contribute to reducing poverty in the area? • What other impacts does this intervention have on the living conditions of the wider community?

The connection between evaluation questions and evaluation criteria. Another way to think about evaluation questions is to think about the common criteria for evaluation as originally defined by the Organization for Economic Cooperation and Development (OECD). As already mentioned, evaluations are periodic assessments of the *relevance*, *efficiency*, *effectiveness*, *impact*, and *sustainability* of our intervention (OECD 1991). Taking a closer look, we realize that *relevance*, *efficiency*, and *effectiveness* primarily relate to normative questions, while *impact* refers to causality. Questions relating to *sustainability* can be either normative (is the intervention likely to be continued after donor funding ends?) or cause-and-effect (are the observed impacts sustainable over time?). None of these is purely descriptive, though normative questions naturally incorporate descriptive ones. Table 4.2 maps each criterion to the corresponding type of evaluation question.

TABLE 4.2 The connection between evaluation criteria and evaluation questions

Criteria	Description	Details	Type of Evaluation Question
Relevance	Do the objectives match the problems or needs that are being addressed?	• To what extent are the objectives of the program still valid? • Are the activities and outputs of the program consistent with the overall attainment of its objectives?	Normative
Efficiency	Is the project delivered in a timely and cost-effective manner?	• Is the program or project implemented in the most efficient way? • What are the costs per output/beneficiary and how do these compare with similar interventions?	Normative
Effectiveness	To what extent does the intervention achieve its objectives?	• To what extent were the intended results achieved? • What are the major factors influencing the achievement or nonachievement of the objectives?	Normative
Impact	What are the positive and negative changes produced by the intervention?	• What are the higher-level outcomes resulting from the program or project? • What real difference has the activity made to the beneficiaries?	Cause-and-effect
Sustainability	Are there lasting benefits after the intervention is completed?	• To what extent do the benefits of a project continue after donor funding ceases? • What are the major factors that influence the achievement or nonachievement of sustainability?	Normative or cause-and-effect

Source: Based on OECD (n.d.)

[Tip]

Make sure you identify the audience for the evaluation and what that audience wants to know. Some evaluations may be demanded within the organization by program staff or management. Donors or policymakers may require others. Internal and external information needs may be different, leading to different evaluation questions. Involving stakeholders in defining and prioritizing your evaluation questions is therefore crucial.

Source: Adapted from Rubio (2011).

Prioritizing our questions. No type of question is a substitute for the other, though normative questions usually include and build on descriptive ones. All are looking at different aspects of the project and provide a different type of information that can be useful. If we want to focus on results, however, then cause-and-effect questions have a special appeal. In fact, if our goals are to identify promising youth livelihood interventions and to prove what effects our intervention really have, then cause-and-effect questions should be a part—if not a priority—of our program's learning objectives.

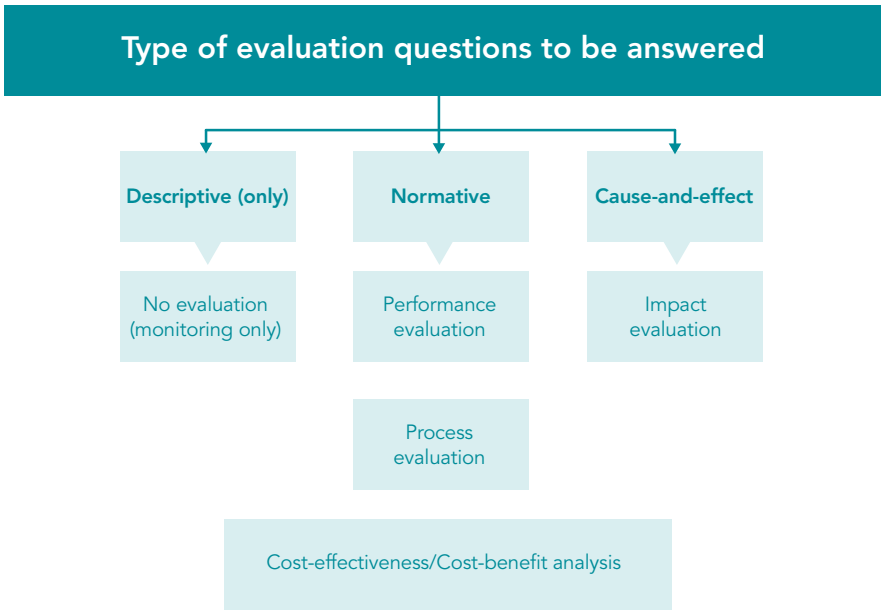
Each of these three kinds of questions—descriptive, normative, and cause-and-effect—leads to different considerations for the type of evaluation to be set up. Program managers and evaluators can allocate a potential question into one of the three types and then consider the implications of each type of question for the development of an evaluation design. Thus, by choosing a set of evaluation questions we define the menu of appropriate monitoring and evaluation tools that will allow answering them (GAO 1991).

Linking Evaluation Questions to Evaluation Design

There is no "one size fits all" evaluation template. Ultimately, the choice of the evaluation should depend on the preceding questions, not our own methodological preferences or those of the internal or external evaluator. This may seem obvious, but it is not always common practice.

Figure 4.1 provides an overview of available evaluation options depending on the type of questions we want to prioritize.[2]

FIGURE 4.1 From evaluation questions to evaluation design



No Evaluation

If a program manager requires only descriptive information about the intervention, for example, because the project is in a very early stage and the objective is to obtain some general information about how the program is being implemented, then a full-fledged evaluation may not be necessary. In that case, the knowledge obtained from monitoring may well be sufficient. Obviously, this requires the existence of a well-functioning monitoring system, with a clearly defined results chain, indicators, data collection tools, and the like (see note 3). If such a system is in place, descriptive information about the program should be available relatively easily.

Performance Evaluation

Performance evaluations assess how well program objectives have been formulated (see criteria in note 2) as well as the program's progress in achieving these objectives (Rubio 2011). They also ask whether the established results framework is appropriate; that is, whether there are inconsistencies among resources, activities, and objectives, and whether priorities or timelines should be adapted to better achieve the agreed objectives. Such evaluations can be carried out across all stages of implementation, but they are particularly common for mid-term reviews (when their focus is on learning for program management) or at program completion (when their focus is on accountability and lessons learned for future interventions). Typically carried out by an independent evaluator, performance evaluations can be implemented relatively quickly and at moderate cost because they rely heavily on desk research and selected interviews.

2 There are other types of evaluations focused on other levels of aid delivery (including sectors, themes, and aid effectiveness) that are not considered in this note. This note is limited to the evaluation of projects and programs.

Sometimes, however, performance evaluations may incorporate more extensive data collection, such as a before-and-after comparison of participant outcomes or additional qualitative tools. While useful for general quality assessment purposes, performance evaluations do not provide absolute certainty about whether the changes observed occurred because of the particular intervention.

Process Evaluation

Unlike performance evaluations, which focus primarily on the achievement of objectives, process evaluations are geared to fully understanding how a program works and seek to assess how well a program is being implemented. They determine whether there are gaps between planned and realized activities and outputs and try to understand the reasons for gaps. Building on descriptive information such as what activities are being offered and who is participating in the program (or who is not), they identify ways to improve the quality of the services offered. A process evaluation may be carried out at specific milestones as an early-warning system or may be conducted when problems such as delays in implementation or beneficiary dissatisfaction have already been detected through standard monitoring procedures (World Bank 2002). Process evaluations tend to rely on a mix of quantitative and qualitative tools, including key informant interviews, user satisfaction surveys, direct observation, and focus groups.

Impact Evaluation

Impact evaluations are the only type of evaluation to specifically answer cause-and-effect questions in a quantifiable manner. Such questions require us to determine not only whether the desired outcomes occurred but also whether those outcomes occurred *because the program was implemented.* As Gertler and colleagues (2011, p. 4) note, this focus on causality and attribution "is the hallmark of impact evaluations" and determines the set of methodologies that can be used. (Note 6 provides an overview of appropriate tools.) To estimate the causal effect of a program on outcomes of interest, any method chosen must estimate the so-called *counterfactual,* that is, what would have happened to program participants in the absence of the program. To do this, impact evaluations require finding a comparison group; that is, a group of people who, in the absence of the intervention, would have had similar outcomes to those of program recipients (Duflo, Glennerster, and Kremer 2006). This is what makes impact evaluations different from all other evaluations. As a result, they tend to require more time and quantitative skills, and they typically cost more than other evaluation types. Based on the information they provide, impact evaluations are particularly useful to inform strategic questions, from scaling up effective interventions to curtailing unpromising programs (Rubio 2011). Moreover, they increase the global knowledge base about the relative effectiveness of different types of livelihood interventions in reaching certain outcomes and help us understand which program design options (dosage, delivery channel, etc.) are most important within a specific program category.

[Definition]

A **counterfactual** refers to the estimated outcomes for program participants in the absence of the program. The counterfactual answers *what would have happened to the beneficiary had the program not taken place.*

Cost-Effectiveness and Cost-Benefit Analyses

Cost-effectiveness and cost-benefit evaluations assess monetary and nonmonetary program costs and compare them with alternative uses of the same resources and the benefits produced by the intervention (Baker 2000). *Cost-effectiveness analysis* (CEA) measures the cost per output or outcome (e.g., $300 per youth trained, $500 per job created) and compares this cost to similar interventions of our own and other

organizations. It thus answers the question about how much output or outcome we get per dollar spent (descriptive) and whether there is a gap with our expectations (normative). *Cost-benefit analysis* (CBA), in turn, weighs the total expected costs against the total expected benefits (outcomes) of an intervention, where both costs and benefits are typically expressed in monetary terms. For instance, if our program were to help 500 youth find and keep jobs or set up sustainable small businesses, we would (1) estimate the aggregate benefits in terms of higher incomes, better health, lower crime, etc., and (2) compare these benefits to the overall costs of the intervention. Since cost-benefit analysis looks at the value of the benefits achieved, it requires a credible estimate of the degree to which the program influenced the outcomes of interest, thereby making it very useful in combination with impact evaluations (for a more detailed description, see note 8). Box 4.1 provides links to examples of the evaluation types discussed above.

BOX 4.1 Examples of evaluation by type

Performance evaluations

- Human Sciences Research Council. 2007. *Mid-term Review of the Expanded Public Works Programme: Synthesis Report*. Pretoria: Southern Africa Labour and Development Research Unit, University of Cape Town; Rutgers School of Law; and ITT (UK). http://www.hsrc.ac.za/research/output/outputDocuments/5465_Hemson_MidtermreviewofEPWPsynthesisreport.pdf
- Education and Employment Alliance. 2010. *An Evaluation of Partnerships in Support of Youth Employability: Global Report*. http://www.iyfnet.org/document/1436

Process evaluations

- Miller, E., and MacGillivray, L. 2002. *Youth Offender Demonstration Project Process Evaluation*. Chapel Hill: Research and Evaluation Associates Inc. http://wdr.doleta.gov/opr/fulltext/YODP_final.pdf
- The Lewin Group, Inc. 2003. *Evaluation Design for the Ticket to Work Program–Preliminary Process Evaluation*. http://www.lewin.com/content/publications/2526.pdf

Impact evaluations

- Attanasio, O., Kugler, A. and Meghir, C. 2009. "Subsidizing Vocational Training for Disadvantaged Youth in Developing Countries: Evidence from a Randomized Trial." IZA Discussion Paper No. 4251. Bonn: IZA. http://ssrn.com/abstract=1426738
- Mensch, B., Grant, M., Sebastian, M., Hewett, P., and Huntington, D. 2004. "The Effect of a Livelihoods Intervention in an Urban Slum in India: Do Vocational Counseling and Training Alter the Attitudes and Behavior of Adolescent Girls?" Working Paper No. 124, New York: The Population Council. http://www.popcouncil.org/pdfs/wp/194.pdf

Cost-effectiveness and cost-benefit analyses

- Elias, V., Nunez, F., Cossa, R., and Bravo, D. 2004. *An Econometric Cost-Benefit Analysis of Argentina's Youth Training Program*. Washington, DC: IDB. http://www.iadb.org/res/publications/pubfiles/pubR-482.pdf
- Jastrzab, J., Masker, J., Blomquist, J., and Orr, L. 1996. *Evaluation of National and Community Service Programs—Impacts of Service: Final Report on the Evaluation of American Conservation and Youth Service Corps*. Bethesda, MD: Abt Associates Inc. http://www.abtassociates.com/reports/ccc_youth_0596.pdf
  (*Note*: This is an impact evaluation and a cost-benefit analysis combined.)

Does Our Operational Context Fit the Desired Type of Evaluation?

As noted by the GAO (1991, p. 15), "It is one thing to agree on which questions have highest priority and to choose an evaluation design. It is quite another to know whether the questions are answerable and, if so, at what costs in terms of money, staff, and time."

After formulating the right questions and identifying a potential type of evaluation, we need to assess the operational context of the intervention to understand what evaluation can be implemented within the given constraints.

Timing

Questions about *what kind* of information is needed are closely related to the question of *when* the results of the evaluation need to be available. Knowing when they need to be available determines when the information needs to be collected.

When is the Demand for Evaluation Identified?

Planning well in advance gives more flexibility in choosing an appropriate evaluation tool. For example, many impact evaluation methods need to be planned even before implementation starts. Planning an evaluation should ideally be part of the program planning (a "prospective evaluation"). In many cases, however, information needs may arise suddenly, for example as a result of sudden problems on the ground, or a request from a donor. Similarly, operational constraints, such as implementing quickly to disburse funds, may dictate the timetable for evaluation. Although these constraints are unavoidable in real life, they reduce the options for evaluation that may be available under such circumstances.

At What Stage of the Program Is the Information Needed?

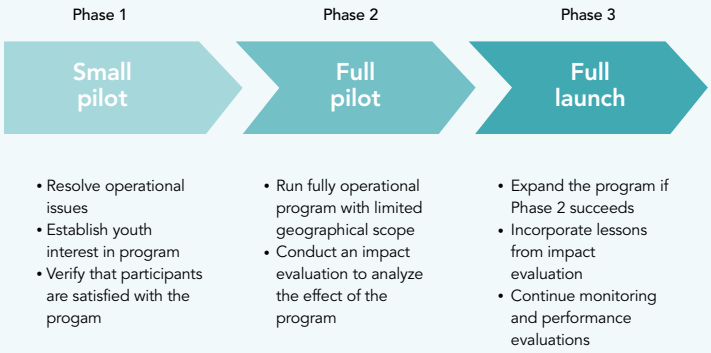
Information needs vary depending on the program lifecycle. For example, a program that has just been planned may require a cost-effectiveness analysis to help determine whether or not to implement the program. Alternatively, for a recently launched intervention, we may need to know how well program procedures are followed and whether any adjustments are necessary to guarantee successful program operation in the future (Rubio 2011). Many times, these information needs can be estimated even before the program begins, and so can the approximate timing of the evaluation.

How Long Does the Evaluation Take?

How long an evaluation takes partly depends on the methods used for collecting and analyzing data, which differ according to the type of evaluation, and on the breadth and depth desired for the particular study, which differ within each type of evaluation. In general, it is fair to assume that pure performance evaluations can be done in one to three months, since they rely heavily on desk research and a limited number of interviews. Process evaluations, in turn, can vary significantly in scope. They may be as fast as performance evaluations, but may take up to six months or longer when complex processes are being analyzed. Impact evaluations tend to be the most time consuming of all (six months to two or more years), since their methodology needs to be well planned and new data collection may be required. Cost-benefit analysis itself can take less than a month if all the necessary data are available. If information first needs to be collected, it can take much longer.

Box 4.2 illustrates at what point in a program's lifecycle different evaluation strategies are best conducted.

BOX 4.2 Lifecycle of a program and suitable evaluation strategies



Source: Adapted from World Bank (2007b, p. 18).

Phase 1. *The first pilot of an innovative and relatively untested youth livelihood intervention is about to start. What evaluation should be used?*

At the earliest stages of a program, we usually need to make sure that everything is being done as planned. Conducting an impact evaluation at this time is not recommended because the results would not reflect the true quality of the program. It is more appropriate to focus on monitoring and process evaluation until the program is fully operational and implementation issues common in setting up new programs are resolved. Qualitative data collection methods (e.g., key informant interviews, focus groups) can be particularly useful in these early stages as they may answer *why* certain elements are or are not working as intended. This initial pilot phase of the program is often referred to as a "feasibility study" to obtain "proof of concept"; that is, to see whether the program can actually be implemented as planned.

Phase 2. *The intervention has been running for one year, and early operational issues have been resolved. Monitoring shows that beneficiaries are satisfied with the program. Should we expand the program or replicate it elsewhere?*

Now may be the time for an impact evaluation. The program is up and running, and we are confident about the quality of implementation. An impact evaluation will allow us to confirm that the program is having an effect on the outcomes of interest. We can also use the impact evaluation to compare the effectiveness of program design alternatives (e.g., different combinations of activities, different intensities of activities) if we are still uncertain about specific design elements. The evaluation will also help us understand some potential unintended effects (positive or negative). As a result of the information obtained through an impact evaluation, we can make the decision on whether substantial funds should be invested in the program or not.

Phase 3. *The impact evaluation yielded very positive results overall. Do we still need to evaluate?*

Although positive results do not imply that the program would work similarly well in different contexts, we can now be fairly confident about the accuracy of our theory of change and the combination of activities. This is a good basis for expanding the program to more participants or replicating it in similar sites. Unless we want to significantly modify our intervention, another impact evaluation will probably not be necessary. However, we need to be certain that the quality of implementation remains high and that we achieve our objectives. Monitoring on all levels, including outcomes, must remain a fundamental component of our program. Moreover, independent performance evaluations in regular intervals can help verify the continued relevance and quality of the program.

Resources

Some otherwise desirable evaluation methods may not be feasible if we don't have the human and financial resources to carry them out. It is important to assess the skills and funding available in our program or organization to ensure they are in line with the needs for the evaluation we envision.

Skills

Conducting quality evaluations requires special skills that may not always exist in a program or organization. In that case, and to ensure neutrality, it is often useful to hire external evaluators. Table 4.3 summarizes some of the major skills required to conduct the various types of evaluations.

TABLE 4.3 Skills required according to type of evaluation

Skill	Description	Performance	Process	Impact	CBA
Program Design and Monitoring	• Familiarity with youth livelihood programming • Experience in program design • General knowledge of quantitative and qualitative data collection techniques • Country knowledge • A university degree in social sciences	!	!	✓	n/a
Quantitative Data Collection	• Specialized training in the design and fielding of surveys • Some knowledge of quantitative data analysis • Program management skills to build and lead a team of enumerators • A university degree in social sciences	✓	✓	!	n/a
Quantitative Data Analysis	• Specialized training in statistics or econometrics • A master's or doctorate degree in economics, public health, or related field	✓	✓	!	✓
Qualitative Data Collection	• Specialized training in implementation of qualitative techniques • A master's or doctorate degree in sociology, anthropology, or psychology	✓	!	✓	n/a
Qualitative Data Analysis	• Specialized training in coding and analyzing qualitative data • A master's or doctorate degree in sociology, anthropology, or psychology	✓	!	✓	n/a
Valuation	• Specialized training in estimating the costs and benefits of human service programs • A master's or doctorate degree in economics, public health, or related field	✓	n/a	✓	!

! *Required;* ✓ *Desirable*

Funding

The differences in scope and varying forms of data collection and analysis create a wide range of evaluation costs. Relying on desk research and key informant interviews is naturally much cheaper than designing and running new surveys with a large number of people. Performance evaluations are therefore usually the cheapest type of evaluation, while impact evaluations tend to be the most expensive (see table 4.4).

TABLE 4.4 Cost estimates for different types of evaluation

Type of Evaluation	Cost	Factors Influencing Cost
Performance Evaluation	$10,000–$30,000	Scope of the evaluation and salary of the evaluator
Process Evaluation	$10,000–$60,000	Same as performance evaluation, but often uses more data collection tools so evaluation can take longer
Impact Evaluation	$15,000–$1 million+	Cost varies widely depending on methodology used: the more data collected, the more expensive the evaluation becomes (see notes 6 and 7 for more details)
Cost-Effectiveness and Cost-Benefit Analyses	$10,000–$30,000	Depends on whether benefits have previously been measured and whether data are readily available

When all data are readily available, impact evaluations can cost as little as $15,000, though in most cases the cost will be above $100,000. Impact evaluations may seem unrealistic for programs with modest budgets. Yet, their cost may be justified if the intervention is—or will be—running for a long time or at large scale. Moreover, the implementing organization does not always have to bear the full cost of an impact evaluation, but can apply for financial assistance to carry out evaluations (see note 7 for more details on budgeting an impact evaluation).

The Political Context

Different stakeholders within and outside our organization may have potentially competing interests in terms of whether or not an evaluation should take place, the issues to be studied, the type of evaluation and its methodology, the data collection strategy, and who, if anyone, should be hired for the evaluation. All of these factors may result in pressures on the choice of an evaluation and influence the relevance and quality of the planned research. Such pressures may range from hints that certain issues should not be studied to an official disapproval from public authorities to interview certain groups of youth, families, or communities.

It is therefore important to try to understand the various interests and the political environment that exists in the specific context. The following questions will help us begin our analysis:

- What are the local political context and the distribution of power?
- What are the relationships among beneficiaries, program managers, policymakers, donors, and other stakeholders?
- What are the interests of and incentives facing each group of stakeholders to influence the conduct of the evaluation and the design of program? For example, if the program is narrowly targeted to one particular group of youth, those not included will have an incentive to influence the program and evaluation in a way that they, too, can receive benefits.
- If the evaluation shows impact, who are the potential winners and losers from any programmatic or policy reform that could derive from the evaluation?
- Will the local environment allow a rigorous and independent evaluation, and will it support the evaluators to publish their evidence-based findings regardless of political consequences?

An international NGO and its local partner in Brazil decided to conduct an impact evaluation on a youth employability-training program they were implementing jointly. After some push and pull, the eligibility requirements were agreed upon, including that the participant selection would be randomized. However, the local partner had a previous agreement with a private corporation that wanted to influence decisions about which youth would be involved in the program, which would bias any evaluation. This conflict made it unfeasible to effectively conduct the study.

Working to understand stakeholder concerns through continuous and open interaction may help us identify ways to address the pressures and competing interests and to build support for the evaluation. Moreover, it is usually helpful to bring in external evaluators who, in addition to contributing a specific skill set, may have an easier time maintaining their independence.

Types of Programs That Usually Justify an Impact Evaluation

Although performance and process evaluations and cost-effectiveness analyses can be part of every program, impact evaluations and cost-benefit analyses should be applied more selectively. According to Gertler and colleagues (2011), the additional effort and resources required for conducting impact evaluations are best mobilized when the program is (1) strategically relevant and influential, (2) innovative or untested, and (3) replicable.

Strategically Relevant and Influential

How important would the results be for informing future programs, policies, or policy dialogue? If the stakes of an intervention are high—for example because a program requires substantial resources and covers, or could be expanded to cover, a large number of people—then an impact evaluation should be considered. This may apply to new initiatives as well as to existing programs when we need to make decisions about their continuation, expansion, or termination. In fact, even an expensive impact evaluation can be highly cost-effective since its findings may help to produce important improvements in program performance. In fact, in the case of large initiatives, even minor improvements may result in considerable savings to the implementing organization (World Bank 2009).

Innovative or Untested

What is the current state of evidence or knowledge on the proposed program's impacts? If little is known about the effectiveness of the type of intervention, globally or in a particular context, an impact evaluation can add powerful knowledge to our organization and the entire field. This is the case for most youth livelihood programs for which the evidence base is still slim (see box 4.3). In the case where no or only little evidence is available, it is usually recommendable to start out with a pilot program that incorporates an impact evaluation. Even if there is existing evidence about a particular type of intervention, an impact evaluation may be still be warranted if the program is implemented in a different context or if it includes innovative aspects that have not been previously tested.

[Online Resource]

Knowledge gaps and potential research questions for impact evaluation

http://www.iyfnet.org/gpye-m&e-resource3

BOX 4.3 Knowledge gaps in youth livelihood programming

Although the following generalizations must be interpreted with caution, we believe existing evidence on youth livelihoods programs appears to be particularly weak in these areas:

Types of programs: Most evaluations exist in the area of training and skills development, while evidence on all other types of interventions such as subsidized employment for youth, employment services, youth entrepreneurship, youth-inclusive financial services, and targeted programs for excluded groups is relatively scarce.

Design Features: Little is known about the relative effectiveness of program alternatives. Within each type of program, what is the effect of adopting different program components, different pedagogies, dosage, and delivery channels?

Context: Evidence of youth livelihood programs is particularly scarce in the Middle East and North Africa, Asia, and sub-Saharan Africa. Moreover, more evidence is needed regarding what interventions and design features are better suited for rural versus urban contexts, informal versus formal settings, or in postconflict and fragile-states environments.

Beneficiaries: How do different types of programs affect young people differently by age group, gender, level of education and socioeconomic background? What works best for disadvantaged groups? And what are the positive or negative spillover effects of livelihood interventions on peers, families, and communities?

Outcomes: What are the effects of livelihood programs not only on employment and labor market outcomes, but also on risky behaviors, civic engagement, family formation, mental health, and the like? Furthermore, evidence on long-term effects of most interventions is virtually inexistent.

For a review of the existing evidence, see the Youth Employment Inventory (www.youth-employment-inventory.org) and Cunningham, Sanchez-Puerta, and Wuermli (2010).

Replicable

To what extent and under what circumstances could a successful pilot or small-scale program be scaled up or replicated with different population groups? If an intervention design is extremely specific and targets a narrow and particular context, then a process evaluation that would contribute to a smooth implementation would probably be sufficient. If, however, the program can be scaled up or can be applied in different settings, then an impact evaluation is an important step in providing the justification for a program to be replicated.

Table 4.5 presents a table summarizing the evaluation types.

TABLE 4.5 Overview of main evaluation types

	Performance Evaluation	Process Evaluation	Impact Evaluation	Cost-Effectiveness and Cost-Benefit Analyses
What are the main questions answered by this type of evaluation?	• Do programs have clear objectives? • Is the program design appropriate to achieve the objectives? • To what extent have program objectives been achieved? • Do priorities need to be changed?	• Are adequate resources and systems (management, information, etc.) in place? • Is the program being implemented according to design? • What are the actual steps and activities involved in delivering a product or service? • What do beneficiaries or other stakeholders know or think about the program?	• How have participants' well-being changed as a result of the intervention? • Are there any unintended consequences, positive or negative, on program participants?	• Are program costs justified compared with similar interventions? • Are aggregate program costs justified in terms benefits achieved?
When can this evaluation be conducted?	It may be conducted at early stages of implementation, for mid-term review, or at program completion	It may be conducted at any time, once or regularly, to confirm that implementation is on the right track or to understand specific operational concerns	It should be designed during the planning of a program, but the final results will typically not be available till after the program (phase) has been completed	It is commonly conducted during an ex ante analysis to determine whether the program is worth implementing or continuing, or after the program is completed to determine the final costs
How long does it take?	1–3 months (more if before/after analysis is included)	1–6 months	• At least 6 months (retrospective evaluation) • 12–24 months (prospective evaluation)	1–3 months
What data collection and analyses are required?	Desk review of existing documents and selected field visits, possibly complemented by monitoring data analysis, beneficiary and stakeholder interviews, mini-surveys, focus groups, etc.	A mix of interviews with program staff and clients, user satisfaction surveys, record review, direct observation, focus groups, and analysis of monitoring data	Statistical and econometric analysis of survey and administrative data, ideally combined with qualitative data analysis	Desk review of existing program documents and relevant literature as well as key informant interviews
Who carries out the evaluation?	Usually independent evaluator (but can also be internal)	Internal or independent evaluator	Independent evaluation team, including lead evaluator, field coordinator, survey firm	Independent evaluator (can be the same as for performance or impact evaluation)
What skills are needed?	Program analysis, possibly simple quantitative methods	Process analysis, quantitative and qualitative methods	Statistical and econometric analysis, possibly qualitative methods	Valuation and economic analysis of program costs and benefits
What are the costs?	$10,000–$30,000	$10,000–$60,000	Cost can range from $15,000 to $1 million or more, depending on the size and complexity of the program	$10,000–$30,000
What programs are best suited for this evaluation?	Every program	Every program	Programs that are: • Innovative and untested • Strategically relevant and influential • Replicable	• Cost effectiveness: Every program • Cost-benefit: Same as impact evaluation

Source: Adapted from Rubio (2011).

Key Points

1. Our learning objectives are the point of departure for any evaluation. This requires formulating evaluation questions across all levels of the results chain and prioritizing the most relevant ones. In general, evaluation questions can be *descriptive*, *normative*, or *cause-and-effect*.
2. The choice of the evaluation strategy depends on the evaluation questions. Purely descriptive information needs may not require an evaluation, and monitoring may suffice. Normative questions are most commonly answered through *process* or *performance evaluations*. If cause-and-effect questions are the priority, *impact evaluations* are needed. *Cost-effectiveness* and *cost-benefit analyses* answer whether the costs involved in an intervention are justifiable.
3. Only impact evaluations—those that can construct a valid counterfactual—allow us to *prove* whether a program has been successful and to generate knowledge that can potentially be generalized beyond the intervention itself. This differentiates them from all other evaluations types and makes them a key instrument for evaluating youth livelihood interventions.
4. Choosing an appropriate type of evaluation depends on the operational context. It is therefore crucial to understand whether the costs in terms of money, staff, and time for each evaluation are appropriate for a given intervention.
5. Since impact evaluations tend to be the most resource intensive type of evaluation, they should be applied selectively to answer strategic questions or to assess innovative pilot interventions testing an unproven, but promising, approach.

NUSAF Case Study: Deciding Whether to Do an IE

Evaluation Questions

The primary learning objective for NUSAF was to estimate the causal impact of participation in vocational training programs on economic livelihoods and social integration. The questions of interest for NUSAF were whether the Youth Opportunities Program helped to:

- increase the number of businesses started
- lower the levels of unemployment
- increase the number of hours working for pay
- improve community integration and decrease conflict
- reduce poverty
- increase psychosocial well-being

Given the cause-and-effect nature of these questions, an impact evaluation was the evaluation method of choice.

NUSAF was also interested in the effects of the program on local training organizations. Since this cannot be easily identified through an impact evaluation, it was decided that this would be part of the monitoring of the Youth Opportunities Program.

Operational Context of NUSAF

Given that NUSAF was a World Bank–funded program with strong support for the impact evaluation from the Government of Uganda, the operational context for an impact evaluation was favorable.

- Timing: The evaluation strategy was planned from the outset of the program. This allowed for the necessary flexibility to plan a rigorous impact evaluation.
- Resources: The necessary resources could be earmarked and a qualified external team hired to conduct the evaluation.
- Political context: Making the evaluation a priority from the beginning fostered stakeholder dialogue and support.

Features of NUSAF that would Justify an IE

The Youth Opportunities Program was a large cash grant program designed and implemented by the government of Uganda. The size and influence of the program, combined with the expectation of rerunning the program in the future, suggested that evaluating the program was an excellent way to increase local and worldwide knowledge of cash grant training programs. Although these types of programs are increasingly implemented, they are generally untested. In addition, the fact that the program was implemented by the government suggested that such a program is scalable and could be replicated in other countries.

Source: Based on Blattman, Fiala, and Martinez (2011).

Key Reading

Imas, L., and Rist, R. 2009. *The Road to Results: Designing and Conducting Effective Development Evaluations,* Washington, DC: The World Bank. (Chapters 6 and 7 are relevant to this note.)
http://books.google.com/books?id=NEsg-BtinIsC&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

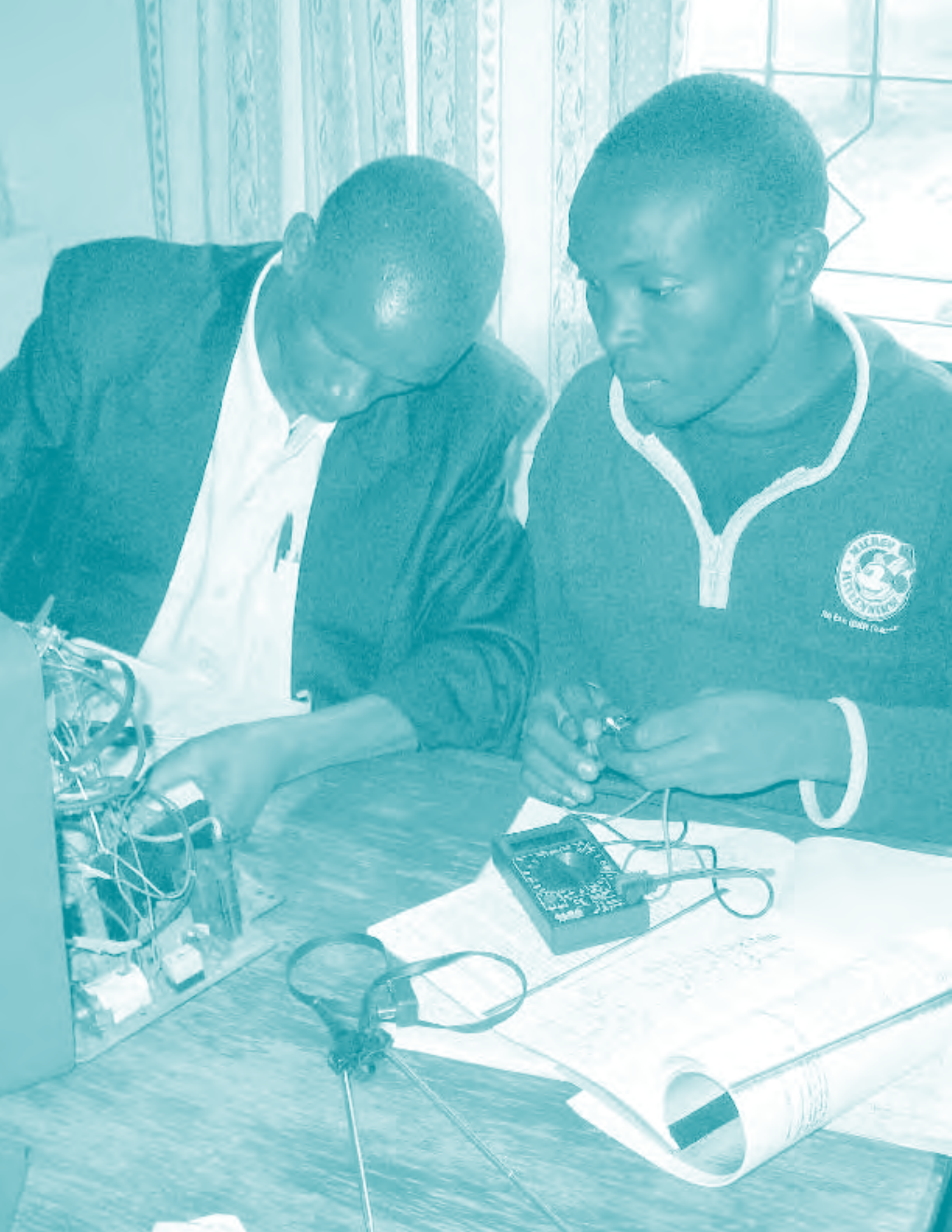
Rubio, G. 2011. "The Design and Implementation of a Menu of Evaluations." PREM Notes, The Nuts and Bolts of M&E Systems, No. 6. Washington, DC: The World Bank.
http://siteresources.worldbank.org/INTPOVERTY/Resources/335642-1276521901256/premnoteME6.pdf

Notes

PART II

ENHANCING PROGRAM LEARNING THROUGH IMPACT EVALUATION

Note: Part II of this guide draws from the extensive field experience by one of the authors as well as on materials originally developed by Duflo, Glennerster, and Kremer (2006), Khandker, Koowal, and Samad (2010), and Gertler et al. (2011), adapting some of the material and illustrations to the youth livelihood field and providing a more concise presentation of impact evaluation methods The authors would like to acknowledge these contributions and recommend Impact Evaluation in Practice *(Gertler et al. 2011) for readers who are interested in a more in-depth description of the contents covered in this second part of the guide.*



NOTE 5: Proving Program Impact

Rigorous skepticism is a creative force
because most damage is done by overconfident people
who thought they knew the answer when they didn't.

— William Easterly

Good intentions are not enough. Instead, we need to know that we are actually improving people's lives and not causing more harm than good without even being aware of it. Proof is provided by impact evaluations, which, unlike other evaluation types, provide scientific evidence of a program's effectiveness.

In this note, we explore the fundamental impact evaluation question: "How can we be sure that the changes in outcomes we see result from our intervention?" We show that measuring impact requires estimating what would have happened in the absence of the program. These estimates can be made by identifying a comparison group through experimental or quasi-experimental evaluation techniques. We also show why the two most common techniques—comparing participants before and after the intervention and comparing participants with subjectively selected nonparticipants—cannot provide reliable estimates of program success.

The Attribution Challenge

Impact evaluations help us answer very specific questions about our program. As discussed in note 4, they try to answer whether an intervention (the cause) improves outcomes among beneficiaries (the effect). For example:

- Does our vocational training program increase trainees' incomes?
- Does our school-based entrepreneurship curriculum increase secondary school completion rates and students' interest in higher education?
- Does our start-up mentoring program foster business creation and sustainability?

Establishing causality between intervention activities and the outcomes we observe can be complicated because other factors may also influence the outcomes we are interested in. For instance, simply observing that business creation increased after our entrepreneurship program was implemented is not proof of our program's success because other factors such as local economic conditions or regulations about starting a business may have improved during the life of our program and contributed to business creation. Similarly, an observed decrease in business creation after our intervention does not necessarily mean that our intervention *caused* a decline in business start-ups; instead it may reflect a worsening of other external conditions.

The purpose of impact evaluations is precisely to overcome this attribution challenge by measuring to what extent a particular program, *and only that program,* contributed to the change in the outcomes of interest.

What Exactly Is "Impact"?

[Definition]

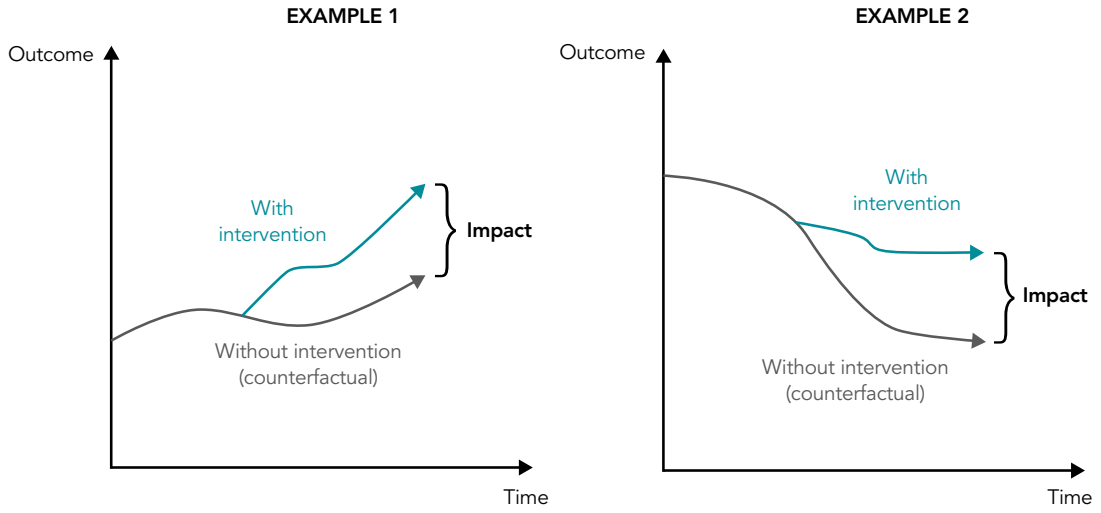
Outcome with the program
– Outcome in the absence of the program
= **Impact**

First, we need to clarify what we mean by *impact*. Often the term refers to higher-level program goals or outcomes relating to changes in overall living standards, such as reducing poverty or increasing the wellbeing of individuals and households. In the context of impact evaluations, however, *impact* is understood more narrowly as the change in outcomes that can be directly attributed to our program. The focus here is on "directly attributed," meaning that we want to know that the changes in outcomes we observe are truly due to our intervention and nothing else.

Simply speaking, as illustrated in figure 5.1, the impact of an intervention is the difference between

- the observed outcomes with the intervention, and
- the observed outcomes for the same individual, household, community, or other unit of observation without the intervention. The outcomes in the absence of the intervention is what we call *counterfactual,* referring to *what would have happened to the beneficiary if the program had not taken place.*

FIGURE 5.1 A visual illustration of program impact



For obvious reasons, it is impossible to observe the same person (household, school, etc.) with and without the intervention. Although we can observe outcomes for those youth that participate in our program, it is impossible to know what their situation would have been in the absence of the program. That is, we cannot know with certainty what would have happened to them if they had not participated in our program. As a result, we will never be able to get the real counterfactual, so an estimate must suffice.

How Can We Estimate the Counterfactual?

To estimate counterfactuals, we identify *comparison groups*, sometimes known as *control groups*. The group of program participants is known as the *treatment group*. A good comparison group has the same characteristics as the treatment group, except for the fact that comparison group members do not benefit from the program.

According to Gertler and colleagues 2011, treatment and comparison groups should share the same characteristics in at least three ways:

1. **They should be identical in terms of *observable* and *unobservable* characteristics.** Observable characteristics refer to age, gender, level of education, socioeconomic status, family characteristics, employment status, and the like. Unobservable characteristics include motivation, interest, preferences, the level of family support, and other factors. Although not every person in the treatment group must be identical to every person in the comparison group, both groups should be the same on average.
2. **Treatment and comparison groups should be expected to react to the program in the same way.** For example, outcomes, such as skills or income, should be as likely to increase for members of the treatment as for those in the comparison group.
3. **Treatment and comparison groups should be equally exposed to other interventions.** For example, both groups should have the same access to other support services provided by local government, NGOs, and so on.

When the above conditions are equal between the groups, then only the existence

[Definition]

A **comparison group** is a group that shares the same characteristics as the group of participants, except for the fact that the people in the comparison group do not benefit from the program. The terms comparison group and control group are often used interchangeably, though strictly speaking the latter is applicable only in the context of experimental evaluations (see below). For the purpose of this document, we will use the generic term *comparison group* throughout.

of the intervention will explain any differences in outcomes. In this case, the causal impact of the program can be demonstrated. If, on the other hand, the comparison group differs from the treatment group in significant ways, we are facing *selection bias*, which will make our impact measures invalid. Selection bias refers to the fact that underlying differences between the treatment and comparison groups by itself explains why we see different outcomes. Selection bias often occurs when the comparison group is made up of individuals who are either ineligible for the program (based on observable characteristics) or who chose not to participate (for unobservable reasons).

[Definition]

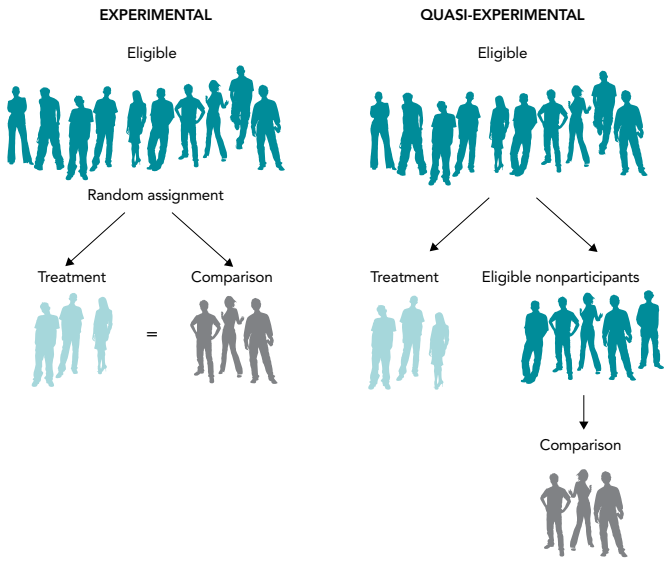
Selection bias usually occurs when program participants and nonparticipants differ in characteristics that cannot be observed, which affect both the individual's decision to participate in the program as well as the outcomes of interest.

In skills training and livelihood programs, it is likely that those who apply to participate are different from those who do not apply, and that these differences cannot be easily seen by the researcher. For example, applicants may be more motivated or have better information than non-applicants. These differences may also mean that applicants, on average, are more successful in the labor market than non-applicants *regardless of the training*. In that case, the better outcomes among training recipients may be due to these underlying differences and not to the training they received in the program.

Techniques to Find Good Comparison Groups

In general, there are two ways to make sure that the treatment and the comparison groups are as similar as possible: (1) with experimental techniques, and (2) with quasi-experimental techniques (see figure 5.2).

FIGURE 5.2 Experimental versus quasi-experimental techniques



Experimental Techniques

Experimental evaluation designs *randomize* who will be in each group. That is, if we have a group of potential beneficiaries (let's say 500 youth, 500 schools, etc.), we randomly select some of them (for example 250) to receive the program. This is the treatment group. The others will not receive the program; this is the comparison group. If randomization is carried out correctly, it is likely that both groups are very similar (1) in observable and unobservable characteristics, (2) in the way they would respond to the program, and (3) in their exposure to other interventions. Evaluations using this

technique, or variations of it, are commonly referred to as randomized controlled trials. See box 5.1 for ethical considerations of randomization.

BOX 5.1 Is randomization ethical?

Some programmers are reluctant to randomly assign potential beneficiaries into treatment and comparison groups. The general concern is that the evaluation leads to withholding seemingly obvious benefits (such as training opportunities) to needy individuals, which would be unethical. In reality, however, it is wrong to assume that one would be denying a benefit if a program has not yet been properly evaluated. In programs that have not been evaluated, random assignment may in fact be more ethical than other selection methods for the following reasons:

- **Uncertainty of program impact.** For most programs, it is not clear if the program has a positive impact on the individual and the community, or if that impact is of a size that justifies the resources being spent. An intervention may in fact have zero impact or even unintended negative side effects. For instance, programs geared toward girls at the exclusion of boys may increase gender violence. A microfinance program for youth may leave participants worse off if they are not able to repay their loans. Even a training program, if designed poorly, may actually decrease job prospects. Where a positive impact is achieved (e.g., a $100 increase in income per participant), it may come at a very high cost (e.g., $1,000 per person), suggesting that the money would be much better spend elsewhere. Thus, in the case of interventions whose impact and cost-benefit structure has not yet been sufficiently proven, it is well justified to evaluate the program based on treatment and comparison groups.
- **Budget constraints.** In reality, because of limited resources, it is rarely possible to serve everyone in need. That is, most programs provide benefits and services only to a limited number of beneficiaries, thereby excluding others, whether this is made explicit or not. For example, if a youth training program has a limited number of available spots, then some youth will receive the training while others will not. Similarly, if an intervention is carried out in one particular district, eligible youth in other districts are excluded. Randomization allows program officers to choose from the universe of potential participants in a way that is fair and that gives the same chance for participation to everyone. If the randomization is done in an open manner (for example as a lottery during a public event), it also enhances transparency in the selection process and may reduce fears in the population that selection was based on personal or political preferences.

It is also important to note that randomized evaluations do not necessarily require denying services to anybody. Note 6 will provide details on different evaluation techniques.

Quasi-Experimental Techniques

Randomization is not always feasible or desirable (see box 5.2). In such cases, quasi-experimental techniques may be used to isolate the effect of our intervention. Although they are usually less reliable than the experimental methods, quasi-experimental designs try to simulate the counterfactual by identifying nonparticipants that are as similar as possible to the treatment group. To do this, quasi-experimental methods usually rely on statistical tools and analysis. Some of the common methods are called discontinuity design, difference-in-difference, and matching (see note 6 for a detailed discussion).

BOX 5.2 Selected examples of when randomization is not possible

- The program has already started; beneficiaries have already been selected.
- Available resources are sufficient to serve all eligible members of the population. It may then be unethical to deny benefits or services only for the purpose of the study.
- We cannot select a comparison group or exclude anyone from the program. For example, a media campaign for financial literacy via TV or radio potentially reaches every household and it is impossible to monitor who listens and who does not.
- The intervention targets a limited number of groups or communities with unique characteristics.
- There is political opposition to providing an intervention to one group and not another.

When the conditions for a good comparison group are met, we say that the impact evaluation has internal validity (see box 5.3).

BOX 5.3 Internal and external validity

Ideally, impact evaluations will satisfy two requirements:

1. They will be *internally valid,* which means we will be able to show causality. To do so, we control for all possible differences between the treatment and comparison group, and are able to clearly attribute changes in outcomes to the intervention. To guarantee this, we use experimental or quasi-experimental techniques (discussed in detail in note 6).

2. They will be *externally valid,* which mean we will be able to generalize findings. That is, we can expect the same results if we provided the program to different or larger groups. To guarantee this, we need an appropriate strategy for choosing the sample of people we work with (this will be discussed in note 7).

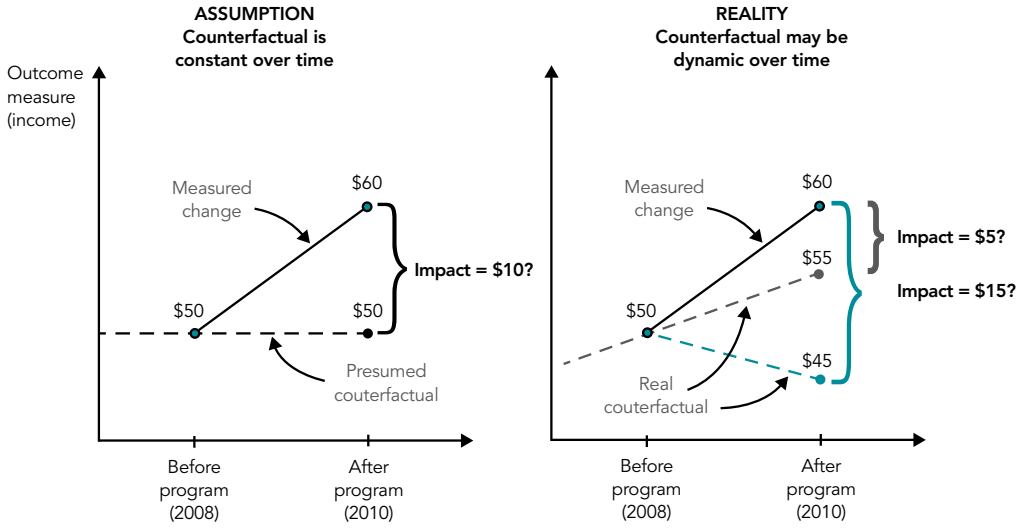
Counterfeit Counterfactuals

The two most common techniques for measuring success in our programs are comparing participants before and after the intervention, and comparing participants with subjectively selected nonparticipants. These techniques fail to identify a quality comparison group. As a result, they cannot be considered proper impact evaluation methods and their impact estimates are usually not credible. Here is why.

Counterfeit Counterfactual 1: Comparing Participants Before and After

In this technique, we use the pre-intervention outcome to estimate the counterfactual. Thus, we assume that if the program had never existed, the outcome for participants after the program would have been exactly the same as before the program. In the example of a training program, we may observe that the monthly income of participants increased from $50 before the program to $60 after the program. We may thus conclude that the impact of the program was $10 per month per person (see figure 5.3, left graph).

FIGURE 5.3 Risks in comparing before-and-after outcomes



The Problem

The assumption that in the absence of the program nothing would have changed is simply unwarranted in most cases. Many things can happen during the implementation period, particularly when programs last several years. For example, local economic conditions may improve, raising the number of available jobs and average incomes; positive weather conditions could raise yields and incomes in agriculture; or the local government could implement its own cash-for-work program, increasing incomes for many youth. If, indeed, the external environment improved independently of the program, then youth would have an increase in income anyway (say, $55 per month), and the real impact of our intervention would likely to be much smaller than estimated by a simple before-and-after comparison. In our example, the gain would be $5 instead of $10 (see figure 5.3, right graph). Conversely, if conditions actually worsened (say youth would earn only $45 in the absence of the program), then we would underestimate the true program impact using a before-and-after comparison.

Conclusion

Many factors can affect the outcomes of youth livelihood interventions over time. As a result, a pre-program outcome measure is almost never a good estimate of the counterfactual. For this reason, a before-and-after comparison is not considered a quality technique to demonstrate impact.

[Tip]

A specific case in which before-and-after comparisons can provide a fairly solid counterfactual is for targeted short-term interventions aimed at improving specific attitudes, knowledge, and skills (see, for instance, the before-and-after evaluation of the ILO *Know About Business* program in box 5.4 below). In that case, the outcomes can sometimes be realistically attributed to a selected intervention. However, other potential program impacts, such as behavior change, employment, and income are influenced by many factors and can thus not be accurately estimated by a simple before-and-after comparison.

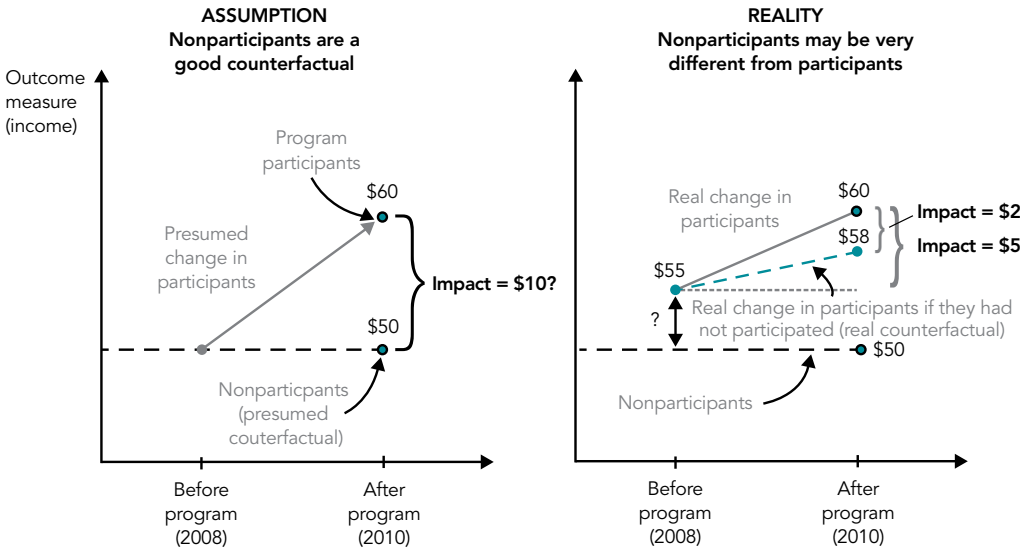
Counterfeit Counterfactual 2: Comparing Participants and Nonparticipants

In this technique, we observe the outcomes of subjectively selected nonparticipants at the end of the intervention to estimate the counterfactual. When comparing participants with nonparticipants, we assume that these groups are very similar in nature. For example, we trust that both groups share the same observable and unobservable characteristics, would react to the program in the same way, and are equally exposed to other interventions.

Using our example of a training program, we would measure the level of income of both participants and nonparticipants at the end of training. Assume we find that

participants earn $60 per month, while nonparticipants earn $50 per month. We then may conclude that our program impact was $10 per month per person (see figure 5.4, left graph).

FIGURE 5.4 Risks in comparing participants with nonparticipants



The Problem

There are two major problems with this approach. First, the assumption that both groups have equal levels of outcome at the beginning of the program may not be true. Participants may have been better or worse off before the program than the subjectively selected nonparticipants. If we measure outcomes only at the end on the program, we may not be able to learn baseline conditions. Participants may already have had a higher income at the beginning of the program than nonparticipants (e.g., $55) and thus the real change compared with our observation at the end of training ($60) would be $5 instead of $10 (see figure 5.4, right graph).

Second, an assumption that participants and nonparticipants are very similar is usually not true. Let's just think about our criteria for selecting young people in the program. Maybe it is on a first come, first served basis. In this case, those with better access to information about the existence of the program, those who live nearby, those who get encouraged by their parents, or simply those who are more motivated to participate would likely end up being part of the program. Alternatively, clear selection criteria such as test scores, interviews, or the quality of a business plan indicate that we explicitly want participants to be different from nonparticipants. In either case, and whether desired or not, participants and nonparticipants are likely to be different from one another on average; therefore, it is misleading to compare the two groups. In reality, given their potentially higher motivation, better access to information, proximity to services, and the like—characteristics that may not always be obvious to us—young people who participated in our program may very well have improved their situation even without the intervention. Going back to our example, if participants would have earned $58 after a certain period even without participating in our program, then their total earnings following the training ($60) would reflect a program impact of only $2, not $10 (see figure 5.4, right graph).

Conclusion

There are usually underlying reasons why some people participate in a program and some don't. These reasons make both participants and nonparticipants fundamentally different from one another, whether we can observe it (test scores) or not (family support, motivation). As a result, subjectively selected nonparticipants almost never represent a good counterfactual to understand how participants would have done in the absence of the program. Therefore, a simple comparison of participants and nonparticipants without using experimental or quasi-experimental techniques is not considered a quality technique to demonstrate impact.

Although the above counterfeit counterfactuals may not be useful to estimate impact—that is, to answer cause-and-effect questions—they may still be of value to our programs. In fact, collecting descriptive information about participants and even nonparticipants over time can be important for program management, since it may help us better understand the dynamics of our program. It is absolutely legitimate to use these types of comparisons as part of our monitoring or performance evaluation, as long as we are aware of what their results can and cannot tell us (see box 5.4 for examples.)

BOX 5.4 Selected examples of non-experimental evaluations

Technique: Before-and-after comparison

ILO Know About Business, Syria

Assessing the Effect of Know About Business (KAB) on the Knowledge and Attitudes of Secondary School Students (2007)
http://www.syriatrust.org/site/images/files/KAB_Schools_Report_0708.pdf

Technique: Comparing participants and nonparticipants

Junior Achievement, USA

The impact on students of participation in JA Worldwide: Selected cumulative and longitudinal findings (2004)
http://www.ja.org/files/long_summary.pdf

Key Points

1. The impact of a program is the change in outcomes that can be directly attributed to the intervention. Understanding impact requires that we isolate the effects of the program from other factors influencing beneficiary outcomes.
2. Measuring program impact requires a counterfactual, knowing what would have happened to our program participants in the absence of the intervention.
3. In order to estimate what would have happened to beneficiaries in the absence of the program, we construct comparison groups that share as many characteristics with the beneficiaries as possible. If a good comparison group can be identified, comparing outcomes between the comparison group and the beneficiaries (treatment group) yields the impact of the program.
4. Impact evaluation techniques to find valid comparison groups can be classified as one of two types. Experimental techniques randomly separate the eligible population into those who receive the program and those who don't. Quasi-experimental techniques try to find a valid comparison group among nonparticipants, mirroring the treatment group as closely as possible.

5. Simple before-and-after comparisons as well as comparing participants with subjectively selected nonparticipants do *not* provide credible impact estimates. The first fails to control for changes in external factors over time, the second fails to control for (often unobservable) characteristics that influence program placement. However, both can be useful for providing descriptive information as part of our monitoring system.

NUSAF Case Study: Identifying a Counterfactual

Identifying the counterfactual was an especially important concern for the Youth Opportunities Program evaluation. Are NUSAF participants different from the general population? If so, how could a counterfactual be drawn from them?

The government and research team expected that there would be important differences between the NUSAF participants and the general population. One clue was that individuals were supposed to form groups and submit proposals. This meant the applicants would need to be at least somewhat educated, implying they are better off. In addition, those who submitted proposals to the program had to want to be engaged in business, so they were probably very motivated. This is not a characteristic that is easily measured.

To verify potential differences, NUSAF looked at the characteristics of program participants collected at baseline and compared them to other youth surveyed around the same time. It was found that Youth Opportunities Program members owned significantly more assets and were much more educated than the general population. Additionally, women were highly underrepresented in the program (33 percent) compared with the general population (51 percent). Households in the study were five times more likely to own a radio or bicycle and three times more likely to own a mobile phone or cattle than the general population. There were also disparities in education among program participants and the general population.

In addition, by comparing rates of poverty in the general population with those of program participants, striking differences were found: at least 50 percent of the participants were above the defined levels of poverty. Thus, whether considering relative or absolute difference between the general population and Youth Opportunities Program participants, it was evident that applicants to the program, on average, were part of higher socioeconomic strata than a representative sample of youth in the region.

The differences across groups underlined why a careful impact evaluation was necessary. Using the general population as a counterfactual would greatly overestimate the effect of the program, as there were already major differences in the sample populations even without the program.

To identify a valid counterfactual, the evaluation team could take advantage of the fact that there was a very high demand for the program, but few remaining funds. The problems with identifying an appropriate comparison group outlined above, along with the lack of funds to ensure everyone who was eligible could participate, led to the decision to use randomized methods in order to select the comparison group.

Source: Blattman, Fiala, and Martinez (2011).

Key Reading

Gertler, P., Martinez, S., Premand, P., Rawlings, L., and Vermeersch, C. 2011. *Impact Evaluation in Practice*. Washington, DC: The World Bank. See Chapter 3. http://www.worldbank.org/ieinpractice

Notes

NOTE 6: Identifying an Appropriate Impact Evaluation Method

The objective of this note is to provide practitioners with an overview of the different tools available for an impact evaluation and to provide guidance on which tool may be the most appropriate for a particular program. We present a toolbox of six methods commonly used in impact evaluation, organized by their ability to construct a counterfactual with minimal bias. Each technique has advantages and disadvantages. The choice of an impact evaluation method will depend not only on the theoretical quality of the method, but also on the operational context of the program. Program managers therefore need to be involved during the evaluation design to make sure the evaluation responds to the needs and context of the intervention.

Choosing Among Impact Evaluation Methods

Every impact evaluation technique differs in terms of the circumstances in which it is best applied; every evaluation does not fit every program context. The characteristics and circumstances of our program will thus guide our selection of the impact evaluation method to be used. In particular, as Gertler and colleagues (2011, pp. 143–149) illustrate, we need to consider timing, coverage, targeting, and resources.

Timing

Has the program already started? The key issue here is whether the impact evaluation can be incorporated into the program design. As will be explained in more detail below, when an impact evaluation is planned from the outset of the program, the quality of the evaluation will be greatly increased and a much larger scope of methodologies can be used.

Coverage

Can the program serve all eligible people? Ideally, we would like to serve every young person in need. This is easier for some types of programs than for others. If the program offering is not resource intensive (such as opening savings accounts for minors) or if it is provided via mass media channels (financial literacy campaign via radio or TV) then we may not want to—or even be able to—exclude anyone from benefiting from the intervention. In most cases, however, we do not have enough resources to provide our youth livelihood programs to everyone who is eligible, forcing us to decide which of the eligible youth will receive the program and which will not. Although not being able to reach every youth may be frustrating from a programming perspective, excess demand offers opportunities to identify a comparison group and conduct quality assessments on the impact of our program.

Targeting

How does our program select beneficiaries? Unless we are able to provide the program to all eligible youth, the selection of individuals or groups occurs by the following means:

1. **Random assignment** is the process of giving each individual or group an equal chance to receive benefits. Drawing names out of a hat to decide who will receive job training now and who will be waitlisted is one example.
2. **Eligibility ranking** determines eligibility according to clear criteria using a cutoff point or threshold. Providing scholarships based on test scores, or providing training based on income levels are examples of eligibility ranking.
3. **Selective targeting decision.** Sometimes there are no clear criteria for why one individual or group is selected over another, which, rather than ensuring fairness in selection, leads to a biased selection of participants. Cases such as first come, first served practices; political factors; and reasons of practicality are examples of inherently subjective selection methods.

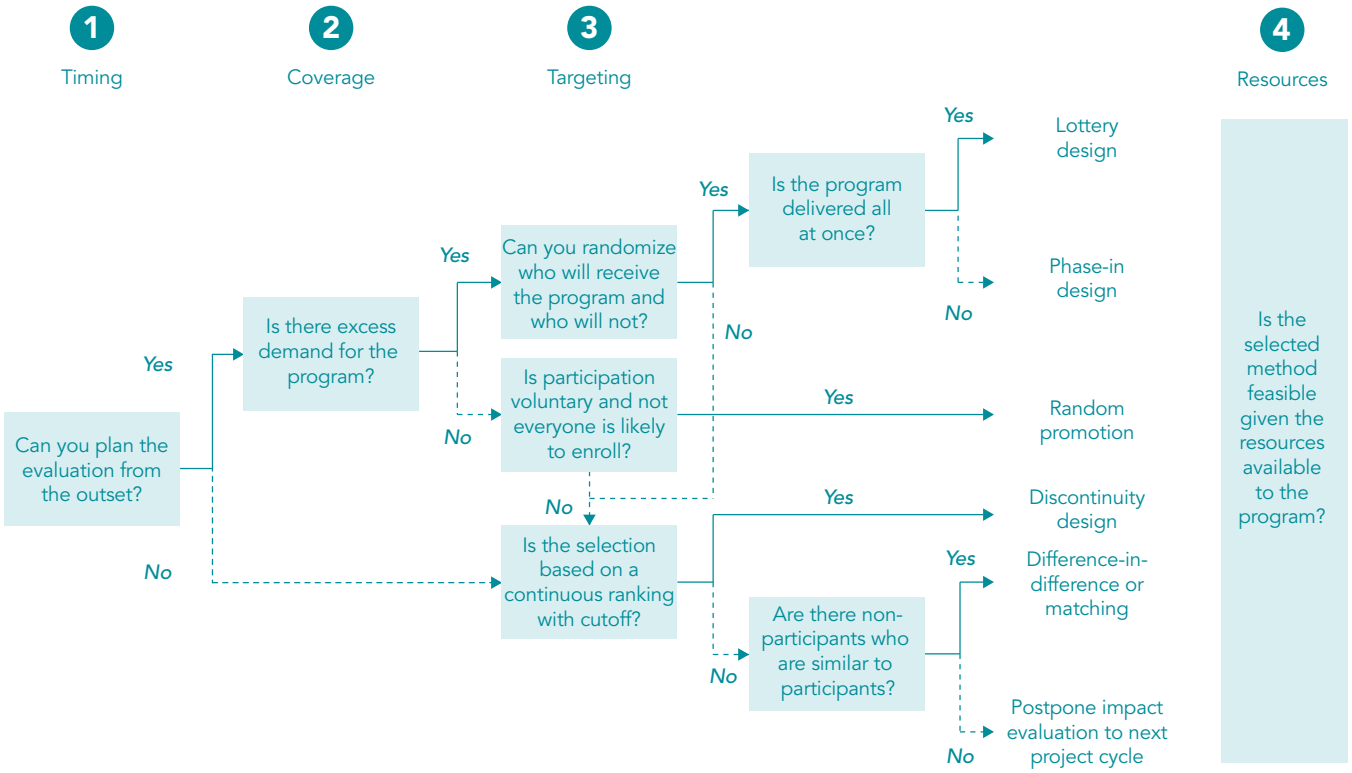
Resources

Does the program have the resources to carry out a specific impact evaluation? Impact evaluation techniques have different requirements in terms of sample size, data collection, complexity of statistical analysis, and cost. Even when we identify a method

that would fit our operational context, it may or may not be feasible given the resources available to us.

The four questions above should be in the back of our minds as we consider various impact evaluation techniques. The answer to these questions will determine which of the six methods is best in our context (see figure 6.1). A discussion follows of the evaluation methods themselves.

FIGURE 6.1 Decision tree for choosing impact evaluation techniques



Sources: Elaborated upon GAO (1991, p. 69); Duflo, Glennerster, and Kremer (2006, pp. 24–27); Gertler et al. (2011, p. 148).

[Definition]

A **randomized controlled trial** is a study in which people are allocated at random (by chance alone) to receive a treatment, such as participating in a specific intervention.

A **sample** is a subset of a population. Since it is usually impossible or impractical to collect information on the entire population of interest, we can instead collect information on a subset of manageable size. If the subset is well chosen, then it is possible to make inferences or extrapolations to the entire population.

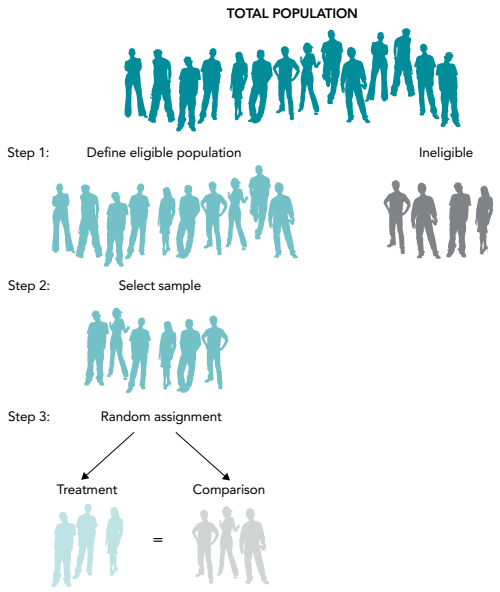
Method 1: Lottery Design

A lottery is a simple and transparent way to assign youth to groups who will receive our services (the treatment group) and those who won't (the comparison group). This is the method used to design randomized controlled trials. It is a statistical regularity that if a large enough sample of people from the same population of interest are randomly assigned to one of two groups, then both groups will, on average, have similar observable characteristics (age, gender, height, level of education, and the like) and unobservable characteristics (such as motivation and state of mind). Through randomization, the difference in outcomes we observe between the two groups at the end of our program can be attributed to the intervention because all other factors that could influence the outcomes are, on average, equal. Lottery designs are considered the most robust type of impact evaluation, so the results are usually the most trusted by donors, stakeholders, and governments.

How It Works

There are three steps to a lottery design (see figure 6.2).

FIGURE 6.2 Steps in a lottery design



Step 1: Define the Eligible Population

The first step in a randomized controlled trial is to find a group of eligible young people for a program. If a medical scientist is studying the effect of a drug on a childhood disease, she searches for a specific group of children and will not enroll adults or elderly people in the program. Likewise, a youth livelihood program may target urban street youth of a specific age range, and so will not include adults or rural youth. What is important here is to have very clear and transparent criteria (age, gender, income level, employment status, etc.) and to be able to communicate who will be eligible to join the program and who won't.

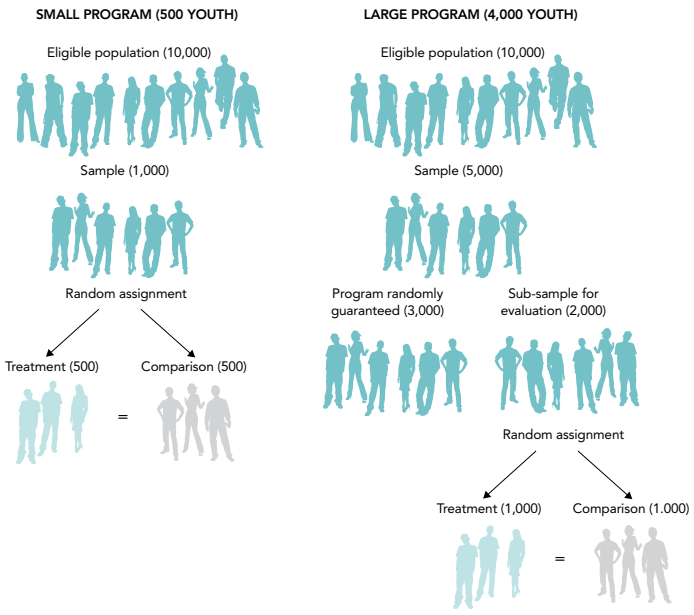
Step 2: Select a Sample for the Evaluation

To evaluate an intervention, we do not need to test everyone who will participate in the intervention. We just need to choose a representative group of people that is numerous enough for the purpose of our evaluation; this is called our *sample* (see note 7 for more details about how to determine the sample and its size). These will be the youth on whom we will collect data.

Choosing the sample for the evaluation can be done in two ways, depending on whether the program is large or small. A small program may find that there are 10,000 eligible beneficiaries, such as urban street youth aged 16–24. The program may have the budget to help 500 of them. Ideally, a comparison group will be equal in size to the treatment group, so 1,000 out of the 10,000 street youth will need to be selected for the program and evaluation (see figure 6.3, left image).

Large programs may be bigger than the sample size needed for an evaluation. If the program is able to serve 4,000 youth, it is not necessary to find an additional 4,000 youth for comparison. Instead, only 1,000 may be needed. The program can then identify a sample of 5,000 youth from the total population of 10,000. Of these, 3,000 youth can be guaranteed admission to the program. The remaining 2,000 will then be randomly split between the program and the comparison group (figure 6.3, right image).

FIGURE 6.3 Choosing samples for small and large programs



In order to make the selection representative of the total eligible population of 10,000 street youth, the sample (whether 1,000 in the first case or 5,000 in the second case) should be selected at random from the eligible population. By selecting randomly, the program participants will, on average, have similar characteristics as the total eligible population. Even though we include only a limited number of youth in the study, the potential impact of the program can be generalized to the entire eligible population, in this case, 10,000 youth.

[Tip]

One way of getting a random sample of youth is to get a list of the total population of street youth from a census, voter registration records, or some other database, and randomly select from that list. If that is not possible, randomly targeting areas where street youth interact, such as an urban center, will produce a random sample. If youth are known to spend time at 50 different centers around a city or country, randomly selecting centers and then selecting a portion of youth at these centers to participate in the study will likely result in a selection of youth with minimal bias. Note 7 will discuss sampling more in detail.

Step 3: Randomize Assignment

The next step is to assign the selected sample of youth to treatment and comparison groups roughly equal in size. In randomized controlled trials, every youth has the same chance of receiving the program. Randomization can be via traditional techniques such as flipping a coin, rolling dice, or drawing names out of a hat. Randomization can be done publicly, if desired, if the sample is relatively small (drawing 2,000 names out of a hat, for example, would not be very practical). Alternatively—and more appropriately if the number of people is large—we can randomize by using computer software, such as MS Excel. Randomization can occur at several levels (see box 6.1). By assigning our sample to treatment or comparison groups randomly, we select participants fairly, and we also develop a good counterfactual: if the sample size is big enough, youth in the treatment group have, on average, the same observable and unobservable characteristics as those in the comparison group.

BOX 6.1 Levels of randomization

Randomization can be conducted at the individual, group, or community level, according to program needs.

Individual level. Individual randomization is best for programs in which outcomes will be measured for each participant. There may be problems with this method, such as spillover, which occurs when individuals in the comparison group receive some of the treatment through informal means. For example, youth who received training or other information through our program may share their knowledge or resources with their friends in the comparison group.

Group level. Individual randomization is not always feasible or desirable. If there is not a list of people's names readily available, or if there is an expectation that people selected for the comparison group may receive the program anyway, then randomizing at a group level may be better. This works particularly well for programs that operate on a group level, targeting schools, vocational training centers, youth centers, and the like. In this case, groups of people are randomized into treatment or comparison cohorts. All individuals in the treatment group would receive the same intervention. Randomization at the group level can help reduce spillover effects and may be easier than randomizing on the individual level. Alternatively, it may also be possible to randomize at the subgroup level, such as classrooms in schools.

Village/community level. Programs may also choose to randomize at the level of villages, neighborhoods, communities, or even districts, when activities are implemented on that level, or when spillover effects are expected to occur beyond the group level. For example, if there are 100 villages in a district of interest and we don't have the resources to work with all of them, we may randomly choose to work with fifty of them, while keeping the other fifty villages as a comparison. All the youth within the respective treatment villages would then be eligible to participate in the program.

(continued)

BOX 6.1 (CONT'D) Levels of randomization

Implementing an intervention at a higher level, and, in turn, randomizing at that level, though it may reduce unwanted spillover effects, can also be problematic for the following reasons:

- The higher the level of randomization, the smaller the number of observations that can be compared with one other. Interviewing a number of people per area can mitigate this problem.
- The size of the evaluation sample increases with the scale of the intervention, which can have implications for the cost of the evaluation.
- Higher level units are more likely to experience different external influences over time, which has implications for the comparability between treatment and comparison group, and thus for the internal validity of the evaluation.

Program managers should therefore find the minimum scale of intervention at which the program can be implemented and randomized.

When Can I Use a Lottery Design?

A randomized lottery evaluation is used when the evaluation is planned in advance of implementation (prospective) and when the program can serve only a fraction of eligible youth. As long as resource constraints prevent the program from serving the entire eligible population, there are no ethical concerns in having a comparison group because a subset of the population will necessarily be left out of the program. In such a situation, comparison groups can be maintained to measure short-, medium-, and long-term impacts of the program (Gertler et al. 2011).

With any prospective evaluation, new data will need to be collected, suggesting cost implications. At a minimum, an endline survey (to be discussed in length in note 7) will be required for youth in both the treatment and comparison groups. In many cases, a baseline survey will be needed, as well. Despite the costs associated with collecting new data, a simple random lottery can be the cheapest option for an evaluation because it may require fewer surveys and lower numbers of respondents.

[Definition]

A **prospective evaluation** is one in which participants will be followed in the future, so these studies must be planned as the program is being designed.

Evaluations that look back on participants in programs that have already been implemented or even ended are called **retrospective evaluations**.

Advantages

- A lottery design is the most robust method for developing a counterfactual because it leads to a very well matched comparison group (relying on fewer assumptions than other methods). It is therefore considered the most credible design to measure impact.
- It is by far the analytically simplest of all evaluation methods. The impact of the program in a random trial is simply the mean difference in outcomes between treatment and comparison groups.
- It allows for communities to be directly involved in the selection process for a fair and transparent allocation of benefits.
- Since it is planned from the outset of the program, it can be designed to measure the average program impact and also to compare the effectiveness of different components, different lengths of programming, and so on.
- It is easy to implement and communicate to program staff.

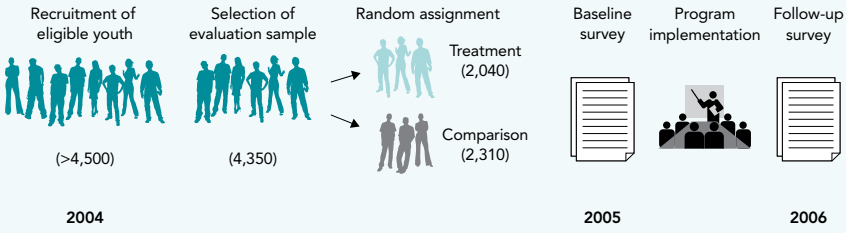
Disadvantages

- It requires a comparison group to be excluded from the program for the duration of the impact evaluation.
- Organizations must ensure that partners and local stakeholders consent to the method.
- The internal validity of a lottery design depends on the fact that the randomization works and is maintained throughout the study, which may not be easy to do. This condition may be threatened if randomization is done incorrectly, if treatment or comparison groups do not comply with their status (that is, if treatment individuals do not take up the program or comparison individuals receive the program), if participants drop out of the study prior to completion, or if there are spillover effects.

Box 6.2 provides an example of a lottery design.

BOX 6.2 Example of a lottery design

Attanasio, Kugler, and Meghir (2009) used a lottery design to study Jóvenes en Acción, a youth employment program in Colombia that provided three months of in-classroom training and three months of on-the-job training to young people aged 18–25 in the lowest socioeconomic strata of the population. The training providers were instructed to recruit more candidates than they had room for in their courses in case not everyone would eventually attend the training. Participants were then selected randomly from the pool of recruited candidates, and the remaining youth were waitlisted and used as the comparison group.



Attanasio and colleagues were concerned that despite randomization, the treatment and comparison groups might be different in ways that the researchers could not control. Using baseline data, they checked the comparability of the two groups and found that, on average, the treatment group had attended school three months longer than the comparison group and had about 5 percent more young women than the comparison group. Neither of these characteristics was thought to significantly influence the treatment outcomes.

The overall results were promising. On average, those who had gone through the program were more likely to be in paid formal employment, have higher incomes, and retain their jobs longer than those in the comparison group. The effects were generally stronger for women than for men.

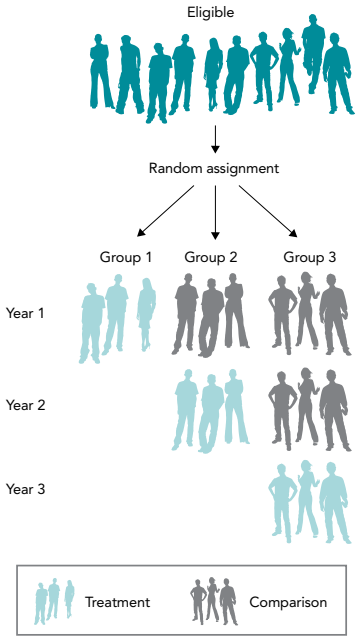
Method 2: Randomized Phase-In Design

Creating a pure comparison group in which youth are never given the program is sometimes impossible. Because many programs are in a community for years, never giving the program to a group of needy individuals can be both politically and programmatically difficult. A variation of the lottery design is the phase-in design. It applies to programs that are rolled out over time, and it uses the natural output flow to develop the treatment and comparison groups.

How It Works

The main difference between a phase-in design and a lottery design is the method of assigning people to treatment and comparison groups. When an intervention is delivered in several tranches over time, a phase-in design gives each eligible person or group the same chance of receiving the program under each of the tranches. One set of youth is then randomly selected to receive the treatment in the first period, while another group is selected to receive the program in the second period, a third group in the third period, and so on. For the time that certain groups are waitlisted, they can serve as the comparison group until they receive the program (see figure 6.4).

FIGURE 6.4 Treatment and comparison groups in phase-in design



Note: Treatment does not necessarily have to stop for the evaluation to work. Some interventions, once in place, will continue to be implemented. However, many programs, such as training, are offered over a limited period of time.

For example, an NGO may have the budget to train 1,500 youths, but it may not have the capacity to conduct all of the training at once. Instead, it chooses to train 500 people per year for three years. If it can identify all 1,500 participants in the beginning, a phased-in randomization may be the best evaluation method for them. The 1,500 youths are randomly split into three groups. In year one, while group 1 receives training, groups 2 and 3 are waitlisted and can serve as the comparison group. In year two, only group 3

remains for comparison. By year three, all three groups will have received training.

As individuals are selected at random for the different groups, it is possible to compare those offered treatment first with those offered treatment later. However, because everyone eventually gets the program, the phase-in design is usually not well suited to finding the long-term impact of a program because eventually there is no comparison group. Even large, longstanding programs will have difficulties asking participants to wait around for three or four years, so the time span of results is often limited to one or two years.

When Can I Use A Phase-in Design?

[Tip]

With a phase-in approach, it is critical to have enough time between each of the phases for the program to show effects. If a program officer believes it will take two years for the impact of the program to take effect, the time between the first and last phase must be at least two years. Small or short-run programs may not be suitable for this approach.

As with a lottery design, a phase-in evaluation is prospective and requires excess demand and the ability to assign participants randomly to treatment and comparison groups. The phase-in design is better suited than a lottery design to large programs that expect to rollout interventions over a number of years. Because the phase-in design requires a set plan for rollout, it also requires a dedicated program team that will be able to follow the rollout through the life of the program.

Phase-in designs do not differ significantly from the lottery design in data or cost requirements. An endline survey will need to be conducted, as well as a baseline survey, in many cases. One important difference is that the program implementation costs may increase because resources will be needed to ensure rollout is implemented in the manner required by the evaluation.

Advantages

- Phase-in designs produce a robust counterfactual, have a fair and transparent selection process, and allow for comparing the impacts of program alternatives.
- The method suits the natural rollout of many programs.
- Because everyone eventually receives the program with this method, phase-in studies can be politically expedient.

Disadvantages

- As with the lottery method, there are challenges to guaranteeing successful randomization and maintaining treatment and comparison groups over time.
- Participants may not wait to join in the program. If they do, there is a risk that they will change their behaviors in the meantime and therefore will not be a comparable comparison group. For example, they may stop looking for jobs in anticipation of joining the program.
- The phase-in method cannot estimate the long-term impact of the program.
- This method requires a clear rollout strategy, which may have operational implications.

See box 6.3 for an example of randomized phase-in design.

BOX 6.3 Example of randomized phase-in design

The World Bank's Economic Empowerment of Adolescent Girls program in Liberia provides six months of training and six months of follow-up activities with two different curricula: (1) skills training for wage employment, combined with job placement assistance; and (2) business development skills combined with links to microfinance. Mentorship is also provided to all beneficiaries starting from the third month of training.

To evaluate its impacts, the World Bank chose a phase-in evaluation design since this would allow for a quality randomized evaluation while also being able to eventually serve all girls who have been promised training. The evaluation took advantage of the natural rollout of the program and the operational constraints that did not allow for training everyone at the same time.

After the baseline survey, 1,273 participants were randomly assigned to the treatment group (receiving training during the Round I of the program in 2010) and 843 to the comparison group (receiving training during the Round II of the program in 2011). The follow-up survey was conducted at the end of each round and complemented with qualitative exit polls to collect information on the participants' views of their training, content, pedagogy, and trainers.

Because the program and evaluation targeted girls who specifically expressed interest in the training, results of the evaluation cannot be generalized to any young woman in the population. The evaluation helps us understand the impact of the training on those who chose to receive training and assistance for wage work or entrepreneurship.

Sources: World Bank (2008); Muzi (2011).

Method 3: Randomized Promotion Design

There may be cases where it is not possible or desirable to exclude any potential beneficiaries either because participation is voluntary and everyone can enroll if they desire or because the program has a sufficient budget to serve the entire eligible youth population immediately. In such cases, the randomized promotion method (also called encouragement design) may be suitable.

How It Works

Randomized promotion identifies the eligible population and chooses a sample just as in lottery or phase-in designs. But it differs in the randomization process. When it is not possible to randomly assign youth into a group that receives benefits and a group that does not, it may be possible to instead randomly promote the program. That is, rather than randomizing those who *receive* the benefits and services, we randomize who is *encouraged to receive* those benefits.

Random promotion is based on the premise that for many programs there will be three sets of potential beneficiaries:

- Youth who never enroll
- Youth who always enroll
- Youth who enroll only if they are encouraged to do so

No matter what the program offers, whether it is free savings accounts, vocational training, or media-based financial literacy programs, it is usually unlikely that every young person who is eligible will want to participate. Some may simply be distrustful of the intervention, others may face constraints such as time or transportation, and others

may just not know about the program.

Random encouragement may take many different forms. In the case of youth savings accounts, we may randomly advertise the initiative in selected schools. For a training program, we could hire a social worker to randomly visit homes of unemployed youth, describe the program, and offer to enroll youth on the spot. In the case of a financial literacy campaign, we may want to randomly send text messages to part of the target audience, but not to others. In all cases, there will still be people in the promoted group that will not take up our program, as there will be people in the non-promoted group who actually will. But the idea is that if the encouragement is effective, then the enrollment rate among the promoted group should be higher than the rate among those who did not receive the promotion. And if the promotion was done randomly, then the promoted and non-promoted groups share, on average, the same characteristics, allowing causal impact to be identified.

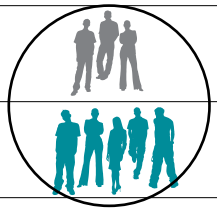
Random promotion evaluation may be suitable for

- *programs that distribute training vouchers.*
- *programs encouraging youth to open saving accounts.*
- *interventions leveraging mass-media based campaigns.*

Unfortunately, we cannot just compare the outcomes of those who participated in the program with the outcomes of those who did not. As discussed in note 5, people who choose to participate in a program are almost always different from those who do not, and many of these differences may not be observable or measurable. Even if promotion is random, participation in the program will not be random, so comparing participants to nonparticipants would be like comparing apples to oranges.

What we can do, though, is compare the outcomes of all those youth who received the promotion with the outcomes of those who did not receive the promotion (see figure 6.5). Let's consider an example of a job-training program in which 30 percent of eligible youth in the non-promoted group and 80 percent of eligible youth in the promoted group participated in the training (Gertler et al. 2011). One year after the program, we observe an average monthly income of $60 for the non-promoted group and $100 for the promoted group.

FIGURE 6.5 Estimating impact under randomized promotion

	Non-promoted Group	Promoted Group	Observed Change
Enrollment (% of eligible population)	30%	80%	50%
Type 1: Never enroll			
Type 2: Always enroll			
Type 3: Enroll only if promoted			
Average outcome (monthly income)	$60	$100	$40
Causal impact			**$80** (=$40/.5)

 *Those who actually enroll in each scenario*

Source: Adapted from Gertler et al. (2011, p. 75).

Given that the promotion is assigned randomly, the promoted and non-promoted groups have, on average, equal characteristics. Thus, the difference that we observe in average outcomes between the two groups ($40) can be attributed to the fact that in the group of people who enroll only if promoted take up the program. Though we cannot directly differentiate them from those who always enroll, we know that their share of the entire population is the difference in enrollment rates (50 percent, or 0.5). Thus, the average impact of the program on those who participated because of the encouragement is 40/0.5=$80.

When Can I Use Random Promotion?

Randomized promotion is well suited for prospective evaluations of programs that have universal eligibility or those in which we cannot control who participates and who does not. It works best when some sort of encouragement can significantly influence take-up. Random promotion is not a good option for services that are extremely popular, such as cash grants, which everyone will want to receive once they hear about it.

With this method, we calculate our average program impact based on people who joined the program as a result of promotional efforts. Because these participants are only a subset of the eligible population, we usually need very large samples for this type of evaluation in order to be sure our results are *statistically significant*. This increases the burden for data collection. If promotion is done on the community level, we may need to survey many more people in the community than we would have had to survey with a simple lottery design. As a result, our costs will likely be higher than costs associated with other types of evaluations. Other conditions are shown in box 6.4.

[Definition]

In statistics, a result is called **statistically significant** if it is unlikely to have occurred by chance. Statistical significance does not tell us anything about the magnitude of the effect size (economic significance); that is, the impact of a program could be statistically significant, yet very small.

BOX 6.4 Necessary conditions for promotion design to produce valid impact estimates

The promoted and non-promoted groups must have comparable characteristics. This can be achieved by randomly assigning outreach or promotion activities to individuals, groups, or communities in the evaluation sample.

The promotion campaign must increase enrollment by those in the promoted group substantially above the rate of the non-promoted group. "Substantially" is a relative concept based on statistical power needs. In general, a program should increase participation by 40 percent or more to be cost effective. This can be verified by checking that enrollment rates are higher in the group that receives the promotion than in the group that does not.

It is important that the promotion itself does not directly affect the outcomes of interest. If the promotion itself changes behavior, it is not possible to determine whether the changes observed in people are due to the program or the promotion. This is most likely to happen if the promotion is done in conjunction with training programs. In most cases, it is most important to know the effect of the program, not of the promotion.

Source: Adapted from Gertler et al. (2011, p. 73).

Advantages

- Randomized promotion campaigns never deny anyone the program, but instead allow people to make their own decisions about whether or not to take up the program.

- This type of evaluation produces a high-quality comparison group with, on average, the same characteristics as the treatment group, just like any of the other randomization methods described above.

Disadvantages

- This method can be used only for specific programs.
- It often needs larger sample sizes than other methods, which increases costs.
- Advanced statistical techniques are required to calculate the program impact.
- Researchers must be careful when interpreting results because the impact estimate is valid only for those who participated in the program because they were encouraged; results cannot be generalized to other groups of potential beneficiaries.

An example of a randomized promotion design is in box 6.5.

BOX 6.5 Example of a randomized promotion design

In South Africa, a randomized promotion design was used to evaluate the impact of entertainment education that aims to enhance the knowledge, attitudes, and behavior regarding sound financial decision making, with a particular focus on managing debt. The program consists of including financial capability storylines in the South African soap opera *Scandal!*, which has been running for several years.

Evaluating the impact of a soap opera on behavior and attitudes is quite challenging. First, it is difficult to separate the effect of the soap opera's message from other messages on similar issues that individuals and families may receive from other sources. Second, certain types of individuals may self-select into watching a particular soap opera, and hence any subsequent behavior change is confounded by these selection attributes. Third, since access to TV is basically universal, it is difficult to establish a good comparison group of individuals who do not receive the financial capability messages.

To overcome these issues, the following randomized promotion methodology was designed: After the study population was identified (approximately 1,000 people), about half the population was provided a financial incentive (about $10) to watch *Scandal!* This was the randomly selected treatment group. Encouragement to watch the program took place through calls before a total of three shows over a period of three months alerting individuals of their financial incentive to watch that particular show. During those calls, treatment group members learned the conditions under which they could receive their incentive and they were asked a number of questions to establish prior knowledge about financial issues. After the show aired, individuals were called and awarded the incentive if they answered several questions about the nonfinancial content of the show correctly. During the same call, they were asked a number of questions on financial knowledge and attitudes.

The same financial incentive was provided for the other half of the population—the randomly selected comparison group—to watch a similar soap opera, one that was aired about the same time and, importantly, did not have a financial literacy component. The mechanism for awarding the incentive was identical to the treatment group. The comparison group was asked the same questions on financial literacy as the treatment group.

The theory was that, if the financial education component of *Scandal!* was successful, those who were encouraged to watch the show would score better on financial questions than the group who was encouraged to watch the other soap opera. Immediate effects on knowledge and attitudes were captured through the short survey after the end of the show; long-term effects were captured through multiple follow-up surveys.

Source: World Bank (2011).

Method 4: Discontinuity Design

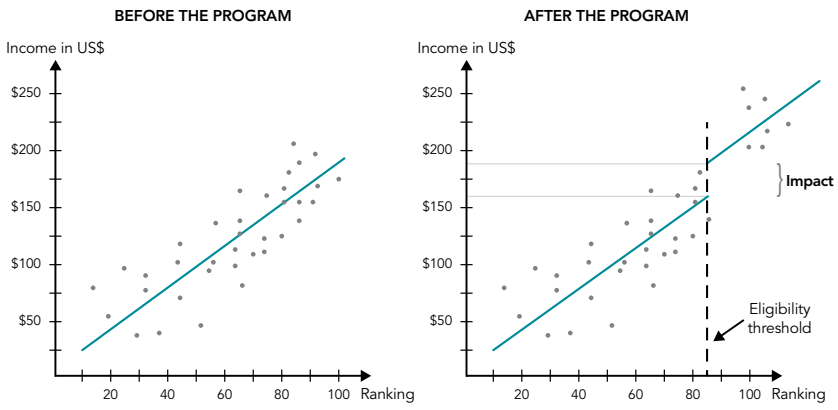
The reality is that in many cases we are not able to plan the evaluation during the program design, and even when we are, it may be impossible to use any form of randomization to obtain a valid counterfactual. In these cases, we may be able to use other targeting rules of the program to obtain a good comparison group. In fact, many programs use a continuous ranking of potential beneficiaries, such as test scores, credit scores, poverty index, or age, and have a cutoff point for acceptance into the program. For example, applicants to a business plan competition or a microfinance bank may be given a score based on a set of criteria and assigned a grade 1–100. If youth score at or above the minimum threshold, say 85 and above, they receive start-up financing. If they score below, they are not accepted into the program. Eligibility rankings like these can be used for an impact evaluation.

How It Works

The premise of discontinuity (or eligibility-index) evaluation designs is that the people who score just above and just below a defined threshold are not very different from one another, or at least the difference may be continuous across the scores. For instance, are applicants who receive a score of 86 much different from those who receive an 84? Probably not. Or are 18-year-olds, who may be eligible for cash-for-work programs, very different from their 17-year-old peers, who may not be eligible? If we have a situation in which some of those youth who receive the program (those just above the threshold) and some of those who don't (those just below the threshold) are not fundamentally different from one another, then comparing the outcomes of these two groups, in turn, would allow us to analyze program impact.

Figure 6.6 illustrates what we may find when analyzing the impact of a youth microcredit initiative. The left graph indicates that, at the time of applying to the program, those who achieved better scores already tended to have higher incomes. There may be many reasons for this, such as that those with somewhat better education are already earning more and that their education also helped them secure better scores. Or those who are more motivated in starting a business were already more entrepreneurial, reflected in higher incomes, and that motivation also helped them convince the jury to support them. Many other explanations are possible, which we do not necessarily need to understand to apply this method.

FIGURE 6.6 Sample discontinuity chart



When starting the program, the local microfinance bank decided that the threshold to receive a loan was 85, and all applicants were accepted or denied support accordingly. Now we'd like to know whether the microcredit program had any impact on incomes. As illustrated in figure 6.6 (right graph), we assume that those who received a score below 85 have the same outcomes as previously, while the income of those with a score of 85 and above increased across the board. From this information, it is possible to identify the impact of the program, which will be represented by the difference in outcomes (that is, the discontinuity of the linear relationship) near the cutoff.

When Can I Use a Discontinuity Design?

The discontinuity design can be used for both prospective and retrospective evaluations. That is, unlike the randomized techniques discussed above, it can also be used when the program is already underway or completed. The main requirement for this method is that program participation is determined by an explicitly specified targeting rule; in other words, by a continuous scale or score. For this method to work, however, we need many observations in the region immediately above and below the cutoff point in order to have sufficient numbers of youth that we can compare with one another. Unless the evaluation is done without baseline data or can take advantage of existing program records, a discontinuity design requires similar data collection as a lottery design, and thus has a similar cost.

Advantages

- The discontinuity method takes advantage of existing targeting rules and does not require any change in program design.
- It provides unbiased estimates for participants near the cutoff.
- It does not require randomization of any kind, so it may be more politically acceptable than other methods.
- It identifies potential effects of marginal scaling. For example, if a program is considering lowering the eligibility threshold from a score of, say, 85 to 75, a discontinuity evaluation can indicate what impact this will have on participants, providing information for a cost-benefit analysis of the proposal.

Disadvantages

- The method requires a very specific threshold for determining groups.
- Impact estimates are valid only for the margin near the cutoff and cannot be generalized to people whose scores are further away from the threshold. The technique does not provide an average impact for program participants.
- It requires large evaluation samples since only the observations around the cutoff can be used.
- As discussed in Duflo, Glennerster, and Kremer (2006), in developing countries, eligibility rules are rarely enforced strictly in the first place, and so there is a high chance that groups may not be distinct, which makes it difficult to obtain valid data using this method.

All in all, the discontinuity method is a good solution when the evaluation starts late or when randomization is not possible. However, it can be applied only in specific circumstances. Box 6.6 presents an example of a discontinuity design.

BOX 6.6 Example of a discontinuity design

Klinger and Schuendeln (2007) use a discontinuity design to study the role of entrepreneurial training on enterprise formation and enterprise outcomes in the context of the business plan competitions run by the NGO TechnoServe in Central America. The program provides training and business development services to help participants prepare a business plan, and it funds a selected number of the best plans.

The evaluators take advantage of the fact that to enter the program there is first a preliminary screening process that assigns applicants a score characterizing their potential entrepreneurial ability. The number of applicants that are admitted into the program is fixed before the competition begins. Applicants are accepted to the workshop if their score falls above the cutoff; if not, they are rejected. This allows for comparing beneficiaries who just received a passing score with those who failed to enter the program by a small margin. Since both groups have similar scores just above and just below the cutoff, it is fair to assume that they also share similar unobservable characteristics, which in turn allows for a high-quality counterfactual.

Statistical analysis confirmed that the eligibility rules were respected—that is, people were selected properly based on their score—and that outcome characteristics of applicants were continuous along their scores prior to the program. After the program, evaluators found a more pronounced change in outcomes around the cutoff. Based on the discontinuity design, in turn, they were able to show that the training increased the probability of opening a business by approximately 10 percent and the probability of expanding a business by more than 20 percent.

Method 5: Difference-in-Difference

In many programs, the selection of target areas and beneficiaries does not follow clear criteria. This can lead to highly selective targeting. For example, we may have prior knowledge about a specific community, better access to some places than to others, or existing partners that already have basic infrastructure in place that we would like to build on. Although there is nothing wrong with this in principle, such subjective targeting rules make it harder to develop a good counterfactual. Nevertheless, we may be able to get a rough estimate of a program's impact by using a difference-in-difference evaluation design.

How It Works

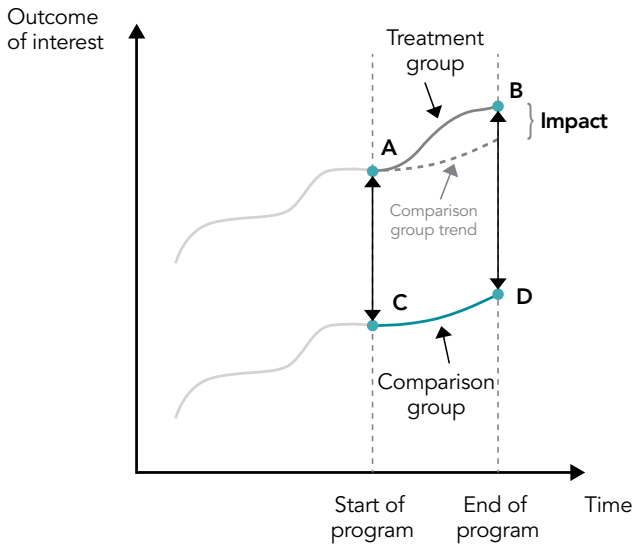
Identifying the comparison group. The difference-in-difference design is basically structured like "a pre-test/post-test randomized experiment, but it lacks its key feature, the random assignment" (Trochim 2006). In the difference-in-difference design, we try to identify a comparison group that we *believe* is similar to our pre-defined treatment group. For example, in center-based youth livelihood interventions, we may pick two comparable training centers or classrooms. In community-based programs, we may use two similar neighborhoods or districts. Either way, we always try to select groups that we think are as similar as possible so we can adequately compare the treated group with the comparison group. However, since the selection is not done at random, we can never be sure the groups are truly comparable—remember that there are unobservable characteristics that we cannot control for—thus, this methodology is also known as the *non-equivalent groups design* (Trochim 2006).

Estimating the impact. As we saw in note 5, simply comparing the outcomes of participants and subjectively selected nonparticipants does not give us the program's

impact, since both groups are most likely different from each other. Similarly, comparing program participants before and after an intervention is problematic as well because many other factors are also likely to influence the participant outcomes over time. But what if we combined both techniques and compared before-and-after changes in outcomes of both a group that enrolled in our program and of a group that did not participate?

Let's imagine a job-training program for youth. To apply the difference-in-difference evaluation technique, we need to measure outcomes (monthly income, for example) for both the treatment and comparison groups before the program begins (see figure 6.7, points A and C) and measure the outcomes of both groups after the program (points B and D). Since both groups are likely to be different from the outset, their incomes at baseline may also be different, but this does not immediately disqualify the method. The difference-in-difference technique compares the difference in outcomes between both groups at the end of the intervention (B minus D) with the difference in outcomes between both groups at the beginning (A minus C). Alternatively, we could compare the difference in outcomes for participants (B minus A) with the difference in outcomes for nonparticipants (D minus C). Subtracting these differences from each other yields a rough idea of the program's impact; it shows whether and how much the training program increased income for participants relative to those who did not participate.

FIGURE 6.7 Example of difference-in-difference analysis



Source: Adapted from Gertler et al. (2011).

[Tip]

A good test for whether it is realistic to assume equal trends between participants and nonparticipants is to compare their changes in outcomes before the program is implemented. If the outcomes moved in tandem before the program started, we can be more confident that their outcomes would continue this trend during the program. If, however, pre-program trends are different, the equal trend assumption may not be correct. Yet, knowing the difference in trends would at least allow us to control for that difference when computing the analysis.

Source: Adapted from Gertler et al. (2011).

The "equal trends" assumption. The underlying assumption of this method is that although the observed and unobserved characteristics of the treatment and comparison groups may be somewhat different (reflected in different levels of income at the beginning), their *differences are constant over time*, or time-invariant. This allows us to use the trend of the comparison group as an estimate for what would have happened to our treatment group in the absence of the intervention.

Is such an assumption realistic? Many observable characteristics, such as year of birth, gender, parent's education, and the like will probably not change over the

course of the evaluation. However, the same cannot be said about several unobservable characteristics, such as personality traits, an individual's intrinsic motivation, risk preferences and so on, which have been shown by numerous studies to change over time, especially in connection with development programs (see, for example, Robins et al. 2001, and Roberts, Caspi, and Moffitt 2003). Therefore, we can never be certain that the differences between the groups do not change over time, which, in turn, could bias our impact estimates. Even if the differences in participant characteristics remained constant, these differences could lead to interaction effects over time. If participating youth are, on average, more motivated than nonparticipants, then they could take better advantage of the program and, in turn, secure higher returns from their participation than nonparticipants would have. Moreover, external factors may influence both groups to a different extent during the implementation period. This would be the case if the municipality starts a new program in our treatment community but not in our comparison community, for example.

When Can I Use a Difference-in-Difference Design?

This design is best used in the absence of a clear targeting mechanism (such as random assignment or eligibility rankings). Since it assumes that the differences of participants and nonparticipants are constant over time, this method is most reasonably used when there are good data at multiple periods before the program begins. There should be at least three data collections: two prior to treatment, and at least one endline. This means that unless the data on participants and nonparticipants are available through other channels, such as an existing household survey, the costs of such an evaluation can be much higher than with other impact evaluation techniques.

Advantages

- The difference-in-difference design provides a way to account for differences between participants and nonparticipants.
- It controls for many individual effects.
- It does not require a prospective evaluation if the necessary data have already been collected.
- It is useful when combined with other methods to increase statistical power.

Disadvantages

- It produces less reliable results than randomized selection methods.
- It cannot be used alone without assuming the treatment and comparison groups change over time in the same way.
- It requires at least three data collections, whereas other methods need only two, so it can be more expensive.

See box 6.7 for an example of this design.

BOX 6.7 Example of a difference-in-difference method

Almeida and Galasso (2008) studied the short-run effects of a program to promote self-employment among workfare beneficiaries in Argentina. Following the severe economic crisis in 2001, the Argentinean government introduced a large workfare program, *Jefes*, including a program initiative to promote self-employment called *Microemprendimientos Productivos* (Productive Microenterprises). The microenterprise program provided in-kind grants to finance inputs and equipment as well as technical assistance through periodic visits of tutors.

To evaluate the impacts of the program in the absence of experimental data, Almeida and Galasso used a difference-in-difference framework. This approach compared the labor market outcomes for program participants before and after the intervention with those of nonparticipants. In order to identify a valid comparison group, they took advantage of the program's promotion campaign, during which *Jefes* beneficiaries could sign up to declare interest in the program. By restricting the comparison group to those who had shown interest in the microenterprise initiative (but eventually did not participate), the authors aimed to minimize the problems of comparing individuals interested in self-employment (for example, due to their entrepreneurial ability or motivation) with those who were not.

A baseline household survey was administered to 309 participants and 244 nonparticipants in November 2004. SIEMPRO, the Argentinean public monitoring and evaluation agency for poverty programs, administered the survey. The same households were re-interviewed one year later, at the end of 2005. With only two data collections available, the evaluators had to assume that in the absence of the program, participants and nonparticipants would have had comparable trends in labor market outcomes (the "equal trends" assumptions).

The findings indicated that, given the relatively low participation rate, jumpstarting self-employment through start-up capital and business training is not necessarily an attractive option for all workfare beneficiaries. Moreover, although the program increased the number of working hours of participants, it failed to significantly increase their average income. Finally, not everyone benefited from the program to the same extent, with positive effects measured only for the more educated participants.

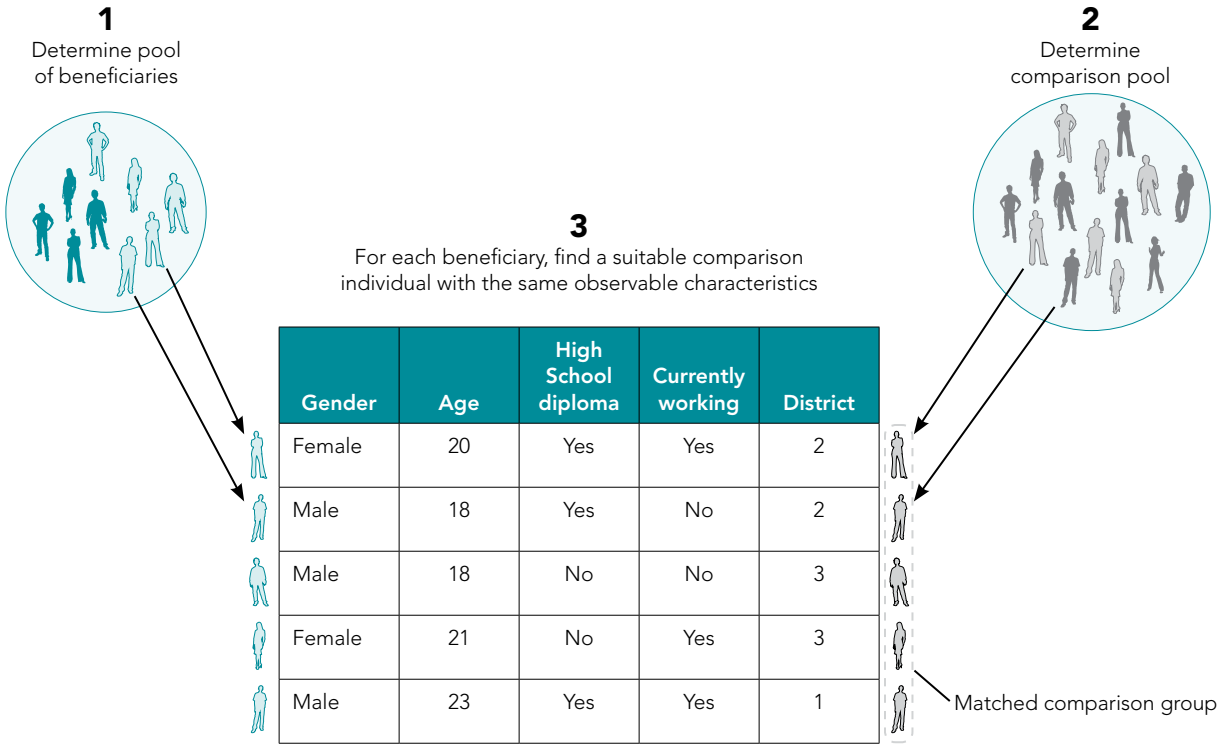
Method 6: Matching

As with the difference-in-difference design, matching is used in the absence of other strict program assignment rules. In the past, matching was popular with program evaluation specialists, but it has become eclipsed by more robust methods, such as those described above.

How It Works

The matching method pairs youth participating in a program with nonparticipants based on observable characteristics (age, gender, level of education, employment status, residency, and other factors). That is, for every individual youth (or group of youths) in the treatment cohort, matching constructs an artificial comparison unit that has as many similar characteristics as possible (see figure 6.8). This statistical technique tries to simulate a comparison group that otherwise does not exist. Ultimately, the average outcomes of those receiving treatment can be compared with the outcomes of the comparison group, and their difference yields the impact of the intervention.

FIGURE 6.8 Exact matching on five characteristics



Identifying a good match for each program participant requires finding those characteristics that explain an individual's decision to enroll in the program. Unfortunately, this is not as easy as it may sound. As Gertler and colleagues (2011) point out, if the list of relevant characteristics is small (as in figure 6.8, above), we will probably find a match for each youth of the treatment group, but each match may not be particularly precise and we run the risk of leaving out other potentially important criteria. If, on the other hand, we want to match based on a large number of characteristics (adding, for example, parents' level of education, test scores, and income level), it may be hard to identify a match for each of the units in the treatment group unless the number of observations in our database of comparison youths is very large.

When Can I Use Matching?

Matching techniques can be used in a variety of settings, regardless of a program's coverage or targeting criteria. In practice, it is often used when none of the other evaluation designs is feasible, especially when the evaluation starts after implementation. Given its inability to control for unobserved characteristics, however, matching is preferably used with one of the other evaluation techniques. Also, in order to match properly, we usually need a large sample size to ensure a matchable comparison group can be found (see box 6.8). If data required have not been collected through other channels, the evaluation may be significantly more costly than other methods described in this note.

[Tip]

The challenge of finding pairs in treatment and comparison groups with many comparable characteristics can be overcome by using a technique called propensity-score matching. Instead of matching treatment and comparison units based on the same characteristics for all selected criteria, propensity-score matching computes the likelihood (the propensity score) of each youth enrolling in the program based on several observed characteristics. Once the propensity score (a number between 0 and 1) has been computed for all participants and nonparticipants for whom data are available, participants are matched with those nonparticipants that have the closest score. These matched nonparticipants then form the comparison group.

BOX 6.8 Steps for applying a matching technique

1. Identify youth that enrolled in the program and that did not.
2. Collect in-depth information on observable characteristics (such as age, gender, level of education, employment status) of enrolled and non-enrolled youth through a baseline survey or by consulting existing data.
3. Using a statistical matching technique such as propensity-score matching, match each participant with a similar nonparticipant.
4. Compare the outcomes of the enrolled youth and their matched comparisons. The difference in outcomes is the impact of the program on that particular individual.
5. Calculate the estimated average impact of the program by taking the mean of the individual impacts.

Advantages

- Matching allows for comparison of outcomes between similar people.
- It can be used with other techniques to validate the quality of the comparison group.

Disadvantages

- Because matching requires direct comparisons of people, a large sample survey may be needed in order to draw an appropriate comparison group.
- Matching can be performed on observable characteristics only. Unobservables, or traits that are very hard to observe, such as personality, motivation, family support, and so on, cannot be incorporated in this technique. It therefore requires an assumption that there are no systemic differences in unobserved characteristics between treatment and comparison groups, which is often implausible. If this assumption does not hold, matching may lead to bias in estimating the impact of the program.
- It may not be possible to find an appropriate match for everyone in the treatment group, impairing the external validity of the impact estimate.

For an example of matching, see box 6.9.

BOX 6.9 Example of matching

Jaramillo and Parodi (2003) used propensity-score matching to evaluate the youth entrepreneurship program implemented by the Peruvian NGO Colectivo Integral de Desarrollo. To estimate the impact of the business plan competition and the subsequent support services consisting of training, follow-up support, and internships on participants, the evaluators constructed a comparison group consisting of those youth who had participated in preparatory activities of the program (pre-training) but either did not join the business plan competition or did not present winning proposals.

The evaluators calculated the probability of an individual's participation in the program based on observable characteristics such as age, gender, level of education, and marital status. Each beneficiary was then matched with someone from the comparison group that had a similar propensity score. The comparison of outcomes (in terms of business sustainability, number of jobs created, income) between the beneficiaries and their matched peers was then used to estimate the impact of the intervention.

However, since the matching could be based only on observable characteristics, there was a realistic chance that the positive effects identified in the evaluation were an overestimate of the actual impact of the intervention. In fact, youth who successfully participated in the business plan competition were likely to be different from youth in the comparison group, for example, in terms of their motivation or skills level, and may have been more successful entrepreneurs than their peers even without participating in the entrepreneurship program.

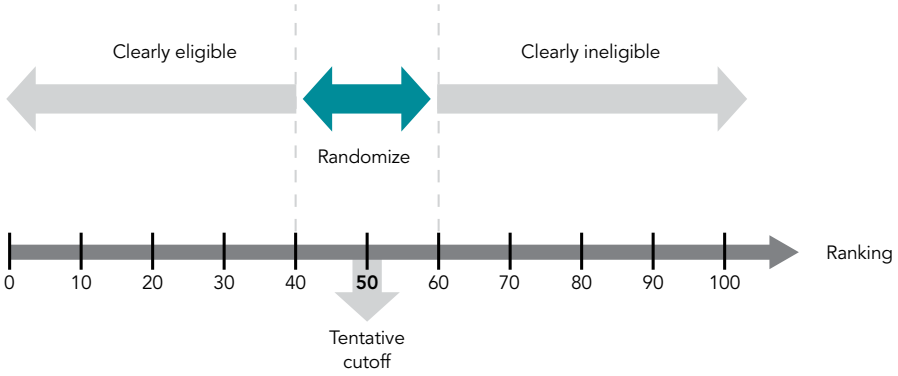
Combining Methods

As we have seen, some methods are stronger in constructing a counterfactual than others. In particular, it may be hard to find good comparison groups when the evaluation is not planned from the beginning of the program. Combining methods may offset some of the weaknesses of a single technique and increase the validity of the estimated counterfactual.

Randomized Discontinuity Design

This technique combines a discontinuity design with randomized assignment. If a cutoff is not clearly designated, or if it is not sufficiently justifiable, it is possible to randomize around the cutoff. In this case, those youths who are clearly eligible are still given the program, while those clearly not eligible are not given the program (see figure 6.9). Only a group near the threshold is selected for randomization. With this method, some of those who otherwise may not have received the program may now receive the program, and vice versa. As in a normal discontinuity design, the results are valid only for those participants at the margin of acceptability. However, given the partial randomization, we can be more confident that the treatment and comparison groups share the relevant characteristics, and we need a smaller sample size to find statistically significant results. The analysis is then done in the same way as any randomized design. The average outcome of those in the treatment group is compared with the average outcome of those in the comparison group, and the difference is the causal impact of the program on those selected.

FIGURE 6.9 Spectrum of eligibility (example of a poverty score ranking)



Note: A lower score represents a higher level of poverty.

Difference-in-Difference or Matching Combined with Randomization

The difference-in-difference technique assumes that those in the treatment and comparison groups are very similar, or at least that their differences are constant over time. Likewise, matching assumes that having similar observable characteristics justifies a comparison between two individuals. Randomization does not require either of these assumptions in order to estimate the impact of a program. However, randomization can be improved when used in conjunction with either or both of these methods. By minimizing differences between those compared, both difference-in-difference and matching methods increase statistical power without the need to increase the number of participants. By combining nonrandom methods with random methods, survey costs can be reduced.

Difference-in-Difference Combined with Matching

If no type of randomization or discontinuity design is feasible, another possibility is to combine the difference-in-difference with the matching technique, thereby mitigating some of the weaknesses both methods have when used on their own. Since the difference-in-difference technique cannot guarantee that treatment and comparison groups are equivalent, combining it with simple matching or propensity-score matching can at least ensure that both groups are very similar in terms of observable characteristics.

For an overview of the standard evaluation methods, see table 6.1.

[Tip]

In practice, the lead evaluator must assess whether it would be useful to combine methods. Practitioners therefore do not need to worry about the details of combined approaches but should be aware that this may be a way to get more reliable impact estimates.

TABLE 6.1 Overview of impact evaluation techniques

Methodology	Description	Comparison Group	Required Assumptions	Data Needed	When to Use?
Lottery	A sample of eligible individuals is randomly assigned into those who receive the intervention and those who do not. Impact is the difference in outcomes between the two groups.	Those selected by lottery	• Randomization is successful and complied with. • The two groups are statistically identical on observed and unobserved factors.	• Post-intervention data for treatment and comparison groups • Baseline data are desirable	• If study can be designed before the program begins • If resources are scarce and it is important to ensure that fair methods are used to enroll needy people into the program • If the comparison group will never get the program for the length of the evaluation
Random Phase-In	Eligible individuals are assigned to treatment tranches and receive the program sequentially.	Those waitlisted	• Randomization is successful and complied with. • Those in the later program phases will not significantly change their behavior while waiting to participate in the program.	• Post-intervention data for treatment and comparison groups • Baseline data are desirable	• If study can be designed before the program begins • If resources are scarce and it is important to ensure that fair methods are used to enroll needy people into the program • If program is rolled out over time
Random Promotion	A random set of individuals or groups is encouraged to enroll in the program. Impact is measured by comparing the average outcomes of those who were encouraged to participate with the outcomes of those who were not.	Those who did not receive the promotion	• Randomization is successful and is complied with. • There will be differential take-up between those who receive promotion and those who do not.	• Post-intervention data for treatment and comparison groups • Baseline data are desirable	• If study can be designed before the program begins • If nobody can be excluded from the program • If participation is voluntary and more people will participate in the program if the program is promoted to them

TABLE 6.1 (CONT'D) Overview of impact evaluation techniques

Methodology	Description	Comparison Group	Required Assumptions	Data Needed	When to Use?
Discontinuity	Individuals are ranked based on specific, measurable criteria. There is a cutoff that determines who is eligible to participate. Outcomes of participants and nonparticipants close to the cutoff line are then compared, and the eligibility criterion is controlled for.	Those close to the cutoff who were not eligible	• Those near either side of the cutoff are very similar in observed and unobserved characteristics.	• Post-intervention data for those nearest the cutoff • Baseline data are desirable	• If randomization is not possible or the evaluation starts after the program begins • If the selection is based on a continuous ranking with cutoff
Difference-in-Difference	Outcomes of program participants and nonparticipants are compared before and after the intervention. The relative change in outcomes is the impact of the program.	Non-equivalent group of individuals who did not participate in the program, but for whom data were collected	Over time, those in the treatment group do not change in a fundamentally different way than those in the comparison group.	• Baseline and follow-up data • Data from at least three time-periods desirable (at least two must be before the program begins)	• If the study starts after the program begins • If nonparticipants who are similar to participants can be identified • If fairness of selection into the program is not considered an issue • Best used in combination with other methods
Matching	Individuals in the treatment group are matched with nonparticipants who have similar observable characteristics.	Exact matching: For each participant, at least one nonparticipant who is identical on selected characteristics Propensity-score matching: nonparticipants who have a mix of characteristics that predicts that they would be as likely to participate as participants	Researchers can identify all of the relevant characteristics of people through a survey.	• Large survey (census, DHS, LFS, etc.), ideally in combination with program-based household survey (ideally two observations of both)	• If the study starts after the program begins • If nonparticipants who are similar to participants can be identified • If fairness of selection into the program is not considered an issue • Best used in combination with other methods

Key Points

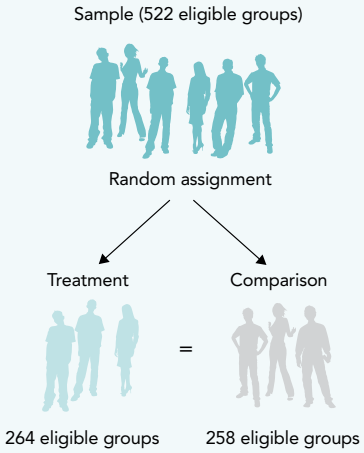
1. Only a selected range of impact evaluation methods allow for obtaining a reliable counterfactual and trustworthy results.
2. Lottery designs, randomized phase-in, randomized promotion, and discontinuity-designs all produce estimates of the counterfactual through explicit program assignment rules. Difference-in-difference and matching methods offer the evaluator additional—though less accurate—tools for impact evaluation when the evaluation starts after implementation and when eligibility criteria are less clearly defined.
3. No single method is best for every program. The best method depends on the operational context (i.e., timing, coverage, and targeting) of the program. Therefore, program managers need to discuss the programmatic constraints with the evaluation specialist because these constraints will affect the feasibility of different evaluation designs.
4. Whenever possible, it is highly desirable to plan the impact evaluation before the program is implemented. Retrospective evaluations tend to be less robust and may not be possible at all if the necessary data was not collected through other channels.
5. In some cases, the methods described here may not be feasible because of budget requirements, timing constraints, or political issues.

NUSAF Case Study: Selecting a Lottery Design

The NUSAF Youth Opportunities Program impact evaluation was developed during program preparation. Because the number of eligible applicants to the program far exceeded the program's funding capacity, the impact evaluation design hinged on the availability of a large pool of eligible but unfunded applications that had been submitted for Youth Opportunities Program funding. Given this large oversubscription to the program, NUSAF management and the program coordinators determined that selection of beneficiaries through a lottery system was not only feasible but also provided a fair and transparent mechanism to allocate funding among equally qualified youth group applicants.

NUSAF District Technical Officers were instructed to verify applications for the minimum set of technical criteria required for eligibility and to conduct field appraisals on programs that would be selected for funding. A list of eligible and verified programs was sent to the Project Management Unit for onward submission to the impact evaluation team, which conducted the lottery for selection of funded proposals. In each district, 30–60 percent of the eligible groups were selected for funding, dependent on budget limitations for that particular district.

Once the complete list of applicants was received from the District Technical Officers, the random assignment of applicants to treatment and comparison groups was completed all at once for each district. Each applicant group was assigned a random number using a random number generator. Groups were then sorted from first to last based on the random number. The sum of the program costs was calculated. Starting from the first randomly selected project, projects were awarded funding until the pools of available resources for that district were exhausted. All other projects remained unfunded and were assigned to the comparison group.



Through this process, a total of 264 projects were selected for funding, comprising the treatment group. The remaining pool of 258 eligible projects not selected for funding made up the comparison group. For the purposes of the impact evaluation, the generation of an equivalent comparison group allowed for the estimation of the counterfactual, the condition that the treatment group would have experienced in the absence of treatment.

Source: Blattman, Fiala, and Martinez (2011).

Key Reading

Duflo, E., Glennerster, R. and Kremer, M. 2006. "Using Randomization in Development Economics Research: A Toolkit." BREAD Working Paper No. 136. (For advanced readers.) http://www.povertyactionlab.org/sites/default/files/documents/Using%20Randomization%20in%20Development%20Economics.pdf.

Gertler, P., Martinez, S., Premand, P., Rawlings, L., and Vermeersch, C. 2011. *Impact Evaluation in Practice*. Washington, DC: The World Bank. (Chapters 4–8 are relevant to this note.) http://www.worldbank.org/ieinpractice

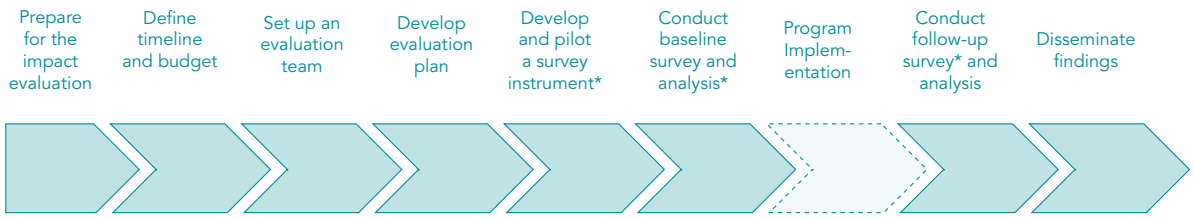
Khandker, S., Koolwal, G., and Samad, H. 2010. *Handbook on Impact Evaluation: Quantitative Methods and Practices*. Washington, DC: The World Bank. (For advanced readers. Chapters 3–7 are relevant to this note.) http://www-wds.worldbank.org/external/default/WDSContentServer/IW3P/IB/2009/12/10/000333037_20091210014322/Rendered/PDF/520990PUB0EPI1101Official0Use0Only1.pdf

Notes

NOTE 7: A Step-By-Step Guide to Impact Evaluation

This note is a step-by-step guide to implementing an impact evaluation for youth livelihood interventions. The information in this note will not replace an impact evaluation specialist, who will always be needed for a proper evaluation. Instead, the note will facilitate planning an impact evaluation from the program perspective, from preparation to the dissemination of evaluation results (see figure 7.1). Moreover, it will clarify the roles and responsibilities of stakeholders involved in the evaluation. We hope to demystify what it means to carry out an impact evaluation and therefore make it easier for each organization or program to consider undertaking an impact evaluation.

FIGURE 7.1 Steps to conducting an impact evaluation



* This step applies only to methods that require data collection by the organization.

Prepare For the Impact Evaluation

Notes 2–6 of this guide clarify the steps that should be taken before initiating an impact evaluation. Ask the following questions:

- **Have I clearly defined my program objective?** The program objective represents what we want to accomplish, the intended result of our intervention. The more concrete the objective in terms of target population, magnitude, and timing of the expected changes, the easier it will be to track progress and carry out an evaluation. For instance: "By 2015, double the income of 1,000 out-of-school youth in Lima, Peru" (see note 2).
- **Have I prepared a results chain?** The results chain provides stakeholders with a logical, plausible outline of how the resources and activities of the program can lead to the desired results and fulfill the program's objective. Every program should put its results chain in writing as it is the basis for monitoring as well as for defining evaluation questions (see note 3).
- **Have I set up a monitoring system with indicators and data collection mechanisms?** Every intervention should have a monitoring system in place before starting an impact evaluation. A monitoring system requires defined indicators and data collection techniques along all levels of the results chain in order to track implementation and results. Without good monitoring in place, the results of an impact evaluation may be of limited usefulness since it will be impossible to determine whether potentially unsatisfying results are due to bad program design or simply bad implementation (see note 3).
- **Have I written down learning objectives and evaluation questions?** Impact evaluation should be based on our information needs. Impact evaluations answer cause-and-effect questions; that is, they determine whether specific program outcomes (usually a subset of those defined in the results chain) are the result of the intervention. Since the type of questions we want answered may vary, we may need to think of other evaluation tools beyond impact evaluation to answer all our questions (see note 4).
- **Have I identified an array of impact evaluation methods?** Before getting started, we should have a basic understanding of the general mechanics of an impact evaluation and the major methodologies that can be used. Knowing the program to be evaluated, we can identify which methodology would best suit our operational context. Having this minimum understanding will help in subsequent discussions with evaluation experts and will facilitate planning (see note 5 and note 6).

In practice, there are often misunderstandings between program managers and impact evaluation experts because the context of the evaluation has not been clearly defined up front. Having a clear idea about how the intervention is intended to work and what should be learned from an evaluation will make the following steps more efficient, saving time and money.

[Tip]

To see whether your program is ready for an impact evaluation and to help you identify an appropriate impact evaluation method, you may want to participate in an impact evaluation workshop in which you can consult with experts about the specifics of your program. Such clinics are offered by the following organizations:

The Youth Employment Network
http://www.ilo.org/public/english/employment/yen/whatwedo/projects/clinics.htm

Abdul Latif Jameel Poverty Action Lab (J-PAL)
http://www.povertyactionlab.org/course

The World Bank
http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTPOVERTY/EXTISPMA/0,,contentMDK:21754074~menuPK:384336~pagePK:148956~piPK:216618~theSitePK:384329,00.html

Define Timeline and Budget

Timeline

By definition, the timing of an impact evaluation is highly dependent on the time frame established by the rest of the program. As discussed in note 6, one of the main questions is whether it is possible to design the evaluation before the start of the intervention, which is always better. It is also important to know when evaluation results are needed. If clear deadlines for obtaining the results exist, for example to inform decisions about program scale-up or policy reforms, we can plan backward from these milestones to see whether we have enough time to conduct the impact evaluation method we are considering.

Some methods require more time to implement than others. Prospective evaluations (evaluations planned in advance), such as all randomized evaluations, naturally have a longer time horizon than retrospective techniques, such as simple matching. Figure 7.2 illustrates the main factors driving the length of an impact evaluation. As we can see, the implementation calendar and the necessary length of time for effects to materialize vary from program to program. As a general rule, prospective evaluations will likely take twelve to eighteen months, and retrospective impact evaluations will take at least six months.

FIGURE 7.2 Sample timeline for a prospective impact evaluation

Task	M1	M2	M3	M4	M5	to	M16*	M17	M18	M19	M20
PROGRAM											
Design program	■										
Identify eligible population		■									
Select participants					■						
Implement program					■	■	■				
Incorporate lessons learned										■	■
M&E											
Design monitoring system**	■										
Develop impact evaluation strategy	■	■									
Set up impact evaluation team	■	■						■			
Develop and pilot survey instrument**		■	■								
Conduct baseline survey**				■							
Analyze baseline data**					■						
Continue to monitor**					■	■	■				
Conduct endline survey**								■			
Analyze endline data									■	■	
Disseminate results											■

* Depends on time needed for effects to materialize

** Applies only to prospective evaluations

In practice, longer lead time for prospective evaluations is less problematic than it may seem. When new programs are set up, they usually take several months to become fully operational. Preparation for the impact evaluation can be done during the program planning and feasibility pilot phases and can easily be ready by the time the program is about to start. Even if a program is already up and running, should the program be organized in phases, a prospective impact evaluation can be planned for the next program phase.

Budget

Impact evaluations can be expensive, which is why many organizations are reluctant to finance them. The reality is that costs vary widely from country to country and across the methodologies and the specific programs evaluated. Evaluations often cost from $100,000 to well over $1 million. In some very specific circumstances, such as when all data are readily available, impact evaluations can cost as little as $15,000. If original data collection is needed, it is unlikely that the design and implementation of an impact evaluation will be less than $50,000.

Cost Drivers

The two major expenses in an impact evaluation are always associated with consultant and staff time and data collection (see table 7.1).

Staff time. The time needed to choose an appropriate evaluation methodology and design should not be discounted. Often the monitoring and evaluation team can design the evaluation in conjunction with an evaluation consultant. The rate of the specialist will range according to experience and can be $200–$1,000 per day, for up to twenty days. More time is needed for data analysis, which can be done by the same consultant who helped design the evaluation. Moreover, additional consultants may be needed to support specific elements of the evaluation, such as survey design. (The next step, Set Up an Evaluation Team, will provide more details about the roles and responsibilities of different evaluation team members.)

Data collection. The main cost component for any impact evaluation is primary data collection. Hiring a survey firm is more expensive than collecting data with program staff but normally ensures better data quality. A benchmark cost per interviewee for a baseline depends on the size of the questionnaire and how easily interviewees can be found. In some cases, a short questionnaire conducted by a survey firm with people that are easily identified with the help of the program staff will cost $20–$40 per interviewee. In places where transport is difficult or where interviewees are not easily found, costs can be $50–$80 per interviewee. This cost includes all aspects of the survey, including hiring and training interviewers, conducting the survey, and presenting the data. Follow-up surveys often present special issues with tracking participants and will likely cost about 1.5 times the baseline. On the other hand, if tracking is not an issue, if the sample population is relatively stable and easy to find, then the follow-up survey may be less expensive than the baseline.

Ways to reduce costs can be found in appendix 2.

Cost Assessment

For most youth livelihood interventions, it is probably fair to assume that the total cost of an impact evaluation will be $150,000–$500,000. This is a lot of money for many small- or mid-sized programs, and it raises the question of whether the cost is justified.

[Online Resource]

List of selected funding opportunities

http://www.iyfnet.org/gpye-m&e-resource4

The evaluation of financial literacy training offered by FINO in India and implemented through local banks is an example of an evaluation that can cost more than the program itself. The pilot program, benefiting about 3,000 participants, cost about $60,000 to implement. The evaluation cost about $200,000. The cost was justified on the basis of scalability. The banking program currently has over 25 million clients in India and is growing by 80,000 people per day. The value of the information from the evaluation is not only for the pilot program but also possibly for millions of future beneficiaries.

Answering this question mainly depends on (1) the time horizon of the program, and (2) current and future funding expectations. For example, if the time horizon for even a relatively small program with an annual budget of $200,000 is five years or more, or if there is potential for scale up to, let's say, $2 million per year, then spending $250,000 on an impact evaluation that informs the design of the larger program is a great use of money. In fact, not conducting an impact evaluation and running an ineffective program would be much more costly. On the other hand, if it is clear that the same program will run for only two years, then the cost of an impact evaluation may be disproportionate, even though the larger youth livelihood community would benefit from the knowledge generated by that study. In such a case, the decision may be made dependent on the availability of external funds to share the costs.

TABLE 7.1 Sample impact evaluation budget

	Design stage				Baseline stage				Follow-up stage			
	Unit	Cost per unit (US$)	No. of units	Total cost (US$)	Unit	Cost per unit (US$)	No. of units	Total cost (US$)	Unit	Cost per unit (US$)	No. of units	Total cost (US$)
A. Staff salaries												
Program Manager	Weeks	2,000	2	4,000	Weeks	2,000	1	2,000	Weeks	2,000	1	2,000
M&E Officer	Weeks	1,000	3	3,000	Weeks	1,000	3	3,000	Weeks	1,000	3	3,000
B. Consultant fees												
Principal investigator	Days	400	10	4,000	Days	400	5	2,000	Days	400	10	4,000
Survey specialist	Days	300	5	1,500	Days	300	0	0	Days	300	5	1,500
Field coordinator/Research assistant					Days	100	80	8,000	Days	100	100	10,000
C. Travel and subsistence												
Staff airfare	Trips	3,000	2	6,000	Trips	3,000	2	6,000	Trips	3,000	2	6,000
Sraff hotel & per diem	Days	150	5	750	Days	150	5	750	Days	150	5	750
Consultant airfare	Trips	3,000	2	6,000	Trips	3,000	2	6,000	Trips	3,000	2	6,000
Consultant hotel & per diem	Days	150	20	3,000	Days	150	20	3,000	Days	150	20	3,000
D. Data collection*												
Surveying					Youth	40	2,000	80,000	Youth	60	2,000	120,000
E. Dissemination												
Report, printing										5,000	1	5,000
Workshop(s)										5,000	1	5,000
Total cost per stage				**28,250**				**110,750**				**166,250**
Total evaluation cost												**305,250**

* Includes training, piloting, survey material, field staff (interviewers, supervisors), transportation, etc.

Source: Adapted from Gertler et al. (2011).

Set Up an Evaluation Team

Impact evaluations require a range of skills, which, in turn, usually requires a big evaluation team. On the one side, there are those responsible for the program, who will determine whether an impact evaluation is needed, formulate evaluation questions, and supervise the overall evaluation effort. On the other side, there are evaluation experts, usually consultants, who are responsible for the technical aspects of the evaluation, including choosing the right methodology, planning data collection, and carrying out the analysis.

The core team consists of the program manager and M&E officer (both internal), a lead evaluation expert (often called the principal investigator, or PI), a research assistant working with the principal investigator, and, for evaluation designs involving new data collection, a survey expert, a field coordinator, and fieldwork team (such as a data collection firm), as well as data managers and processors. Table 7.2 presents the roles and responsibilities of each person. Depending on the size of the program and evaluation and the skill level of the team members, multiple tasks can be assigned to one person.

[Online Resource]

Terms of reference for key impact evaluation staff

http://www.iyfnet.org/gpye-m&e-resource10

NOTE 7

TABLE 7.2 Impact evaluation team and responsibilities

Who	Major Tasks	Profile/Skills Required
Program Manager	• Define learning objectives • Estimate resource requirements • Prepare terms of reference for PI • Hire evaluation consultants	• Experience with designing and implementing youth livelihoods programs • Experience with managing a team • Able to develop budgets • Able to work closely with program and evaluation teams
Internal M&E Officer/Unit	• Define program theory model (results chain) • Define indicators and measurement tools • Manage the monitoring system once the program begins	• Undergraduate or graduate degree in economics, public policy, or related field • Able to work closely with program and evaluation teams • Able to multitask monitoring and impact evaluation responsibilities
Principal Investigator (local or international university, think tank, specialized consultancy)	• Select evaluation design • Adapt theoretically sound designs to real-world budget, time, data, and political constraints • Develop mixed-method approaches • Identify evaluation team and prepare terms of reference • Supervise staff • Determine sampling and power requirements • Analyze data and write report	• Graduate degree in economics, public policy, or related field • Knowledge of the program or similar types of programs • Experience in research methods and econometric analysis • Some experience in the country or region • Demonstrated ability to work effectively in multi-disciplinary teams • Superior written and oral communications skills
Survey Expert (may be same person as the PI)	• Design survey instrument • Prepare accompanying manuals and codebooks • Train the data collection firm • Support piloting and revision of questionnaires	• Graduate degree in economics, public policy, or related field • Experience in surveying children and youth • Experience in carrying out field work in the country or region of interest • Ability to interact effectively with research and program counterparts
Field Coordinator and Fieldwork Team	• Assist in the development of the questionnaire • Hire and train interviewers • Form and schedule fieldwork teams • Oversee data collection • Clean the data so it can be shared with the evaluation specialist	• Legal status, business licenses recognized by the government of the country where work is to be performed • Good network of experienced interviewers, supervisors, and data-entry clerks • Demonstrated 5+ years' experience with organizing surveys on the scale of this program • Strong capacity and experience in planning and organizing survey logistics • Strong capacity in data management and statistics • Ability to travel and work in difficult conditions
Research Assistant	• Analyze data • Support the PI in writing the evaluation reports	• Undergraduate or graduate degree in economics, public policy, or related field
Data Managers and Processors	• Clean the data so the research assistant and PI can use it • Manage data team	• Experience with data software and management of data team

After the initial evaluation design and baseline data collection, and once the program begins, there will be little direct work for the program manager and the M&E officer. It is a good idea to keep one of them, perhaps the M&E officer, working on the evaluation part time during this period to ensure proper monitoring of the program. If

there are any major issues related to the implementation of the program, this will need to be documented and in some cases reported to the larger team.

Not all outside experts should be hired at the same time. The first priority is to select the principal investigator, who should be retained for the entirety of the evaluation, from designing the evaluation to writing the final report, to ensure continuity (though he or she will likely not be working on the evaluation during the implementation of the program). Together with the lead evaluator, other external team members can be selected when necessary. For instance, the survey development expert is normally contracted for short tasks and may be involved in the evaluation for only a few weeks, depending on the size of the evaluation. The data collection firm is hired to conduct the baseline and endline surveys and is ideally the same firm for both data collections, though this is not always necessary or feasible.

Develop an Evaluation Plan

Once the principal investigator is on board, he or she will usually prepare an impact evaluation plan (also called a concept note) in coordination with program leaders. That plan will describe the objectives, design, sampling, and data collection strategies for the evaluation. In essence, the impact evaluation plan (see sample outline in box 7.1) will be the basis for the impact evaluation methodology to be chosen and will guide all subsequent steps in the implementation process of the evaluation.

BOX 7.1 Outline of an impact evaluation plan

1. Introduction
2. Background
3. The intervention
4. The evaluation design
   - 4.1 Objective of the evaluation
   - 4.2 Hypotheses and research questions
   - 4.3 Evaluation methodology
5. Sampling strategy and power
6. Data collection plan
7. Data analysis plan
   - 7.1 Measuring impacts
   - 7.2 Examining differential treatment effects
   - 7.3 Measuring the return of the program (cost-benefit analysis)
8. Risks and proposed mitigation
9. Audience and dissemination
10. Timeline and activities
11. Budget
12. Annexes

[Online Resource]

Resources for finding impact evaluation experts

http://www.iyfnet.org/gpye-m&e-resource5

[Tip]

Partnering with academic institutions is often a powerful strategy for NGOs and governments to develop their impact evaluation capacities. For example

- Save the Children is partnering with Universidad de los Andes in Colombia to evaluate the YouthSave initiative.
- Youth Business International and BRAC are partnering with the London School of Economics.
- The Turkish Ministry of Labor is partnering with the Middle East Technical University on the evaluation of the Turkish Public Employment Agency (ISKUR).

The example of a planned impact evaluation of youth microfinance in Yemen shows the importance of program staff and evaluators collaborating closely from the beginning of a program in order to have a mutual understanding of the operational context. In this case, evaluators independently designed a randomized control trial to assess the impact of lending and other financial services for youth on employment creation, business expansion, and other outcomes. When it came to presenting the evaluation design, the CEO of the bank made it very clear that such a design would be unacceptable in the context of a recently founded financial institution that cannot afford to exclude potential clients for the purpose of an evaluation. The evaluation team then had to start over and finally chose a randomized promotion evaluation design that was more suitable for an intervention with universal coverage.

Developing the evaluation design (point 4) should not be done by the evaluation expert in isolation; instead, the process should closely involve the program staff to make sure the evaluation method fits the learning objectives and operational context of the program (see note 6 for a detailed discussion). In addition, although the principle investigator will certainly approach the program staff and make suggestions for defining the sample for the evaluation (point 5) and planning data collection (point 6), it is still useful for the implementing organization to have a basic understanding of how these aspects are relevant to the evaluation and the program itself. Therefore, we explore these two points in more detail below.

Defining the Sample For the Evaluation

We do not necessarily need to assess every program participant to evaluate an intervention. We just need to choose a representative group of people—a sample—that is big enough for the purpose of our evaluation. If our sample is representative of all eligible youth, we can generalize the results of the evaluation to the total eligible population. That is, we want the results to have external validity, in addition to the internal validity from constructing a good comparison group. To obtain a representative sample, we need a sampling strategy.

We also want the sample to be big enough to be able to generate a reliable comparison of outcomes between those in the treatment group and those in the comparison group. If the sample is too small, we may not be able to see a statistically significant impact of the program, even if there were one. To know how big is big enough we need power calculations. These concepts are discussed below.

[Definition]

External Validity: Our ability to generalize findings. It refers to the extent that we can expect the same results if we provided the program to different or larger groups. To guarantee this, we need an appropriate strategy for choosing the sample of people we work with.

Sampling Frame: The most comprehensive list of units in the population of interest that we can possibly obtain. Drawing from this list allows us to obtain the sample.

Sample: A sample is a subset of a population. Since it is usually impossible or impractical to collect information on the entire population of interest, we can instead collect information on a subset of manageable size. If the subset is well chosen, then it is possible to make inferences or extrapolations to the entire population.

Create a Sampling Strategy

A sampling strategy involves the following three steps:

1. **Determine the population of interest.** First, we need to have a very clear idea about whom we want to target and who will be eligible for the program. For example, age, gender, income level, employment status, and location could determine eligibility. Those who are not eligible will not be included in the study.
2. **Identify a sampling frame.** A sampling frame is the most comprehensive list of units in the population of interest that we can possibly obtain. It tells us how our sample relates to the general population of interest for which we want to we want to extract the lessons of the evaluation. Ideally, then, the sampling frame exactly corresponds to the population of interest, indicating that it would be fully representative. In practice, we would try to get a list of eligible youth from a population census, school or voter registration, or city registry that includes as many of the eligible youth as possible. In reality, however, it is not always easy to obtain a sampling frame that would fully cover the eligible population.
3. **Draw the desired number of units from the sampling frame using one of the available sampling methods.** Various methods can be used to draw samples from our frame, but the most commonly used are some form of probability sampling. With this method, participants are selected into the sample with a specific probability. In the case of random sampling, for instance, every participant in the sampling frame would have the same probability of being included. When non-probability sampling procedures are used, then we are running the risk of creating a sample that is not representative of the eligible population at large.

When we don't have a comprehensive list and don't know how our study population represents the general population of interest, we should not generalize lessons learned beyond the study population. It is tempting to draw general lessons beyond the sample population, and many studies do, but we must be modest and careful when interpreting the results. Similar caution about generalizing conclusions is needed when a program is scaled up, since a larger program may reach youth who are different from those who took part in the original study.

Power Calculations, or "How Big Does My Sample Need to Be?"

It is crucial to know the ideal size of our sample, that is, how many individuals we should draw from the sample frame. If our sample is too small, statistical analysis may lead us to conclude that our program has no positive impact on our beneficiaries, when in reality it does. Conversely, collecting more data than necessary would be very costly. Power calculations help us find the right size by indicating the smallest sample with which it is still possible to measure the impact of our program with a reasonable level of confidence.

Although appropriate sample sizes for evaluations vary, in general, we should estimate having 1,000–3,000 youth in our evaluation to ensure we have enough youth in both the treatment and comparison groups. In some very specific cases, a sample size of fewer than 1,000 youth may be fine. It is almost never advisable to have fewer than 500 participants (250 in the treatment group and 250 for comparison). Evaluation professionals will be able to calculate the appropriate sample size for your particular evaluation.

[Definition]

Power is the probability of detecting an impact if one has occurred. There is always a risk that we will not detect an impact even if it exists. However, if the risk of not detecting an existing impact is very low, we say that the study is sufficiently powered.

[Online Resource]

Example of sample size estimation

http://www.iyfnet.org/gpye-m&e-resource6

[Tip]

Since people may drop out of the program during implementation and hence drop out of the evaluation, it is wise to choose a sample size bigger that the minimum sample indicated by the power calculation.

Planning the Data Collection

The evaluation plan will need to establish the basic data collection strategy. Data collection can be a very complicated task that is best handled by a team of outside experts. Key issues include the timing of data collection, whether new data must be collected, who is going to collect the data, and how the data will be managed. These issues are discussed below.

Timing of Data Collection

The timing of data collection is very important and depends on the nature of the program. When a baseline survey will be used, it should be completed before the program starts and *before participants know if they are going to be enrolled in the program* to ensure their answers are consistent across the treatment and comparison groups. This is critical, as youth may give different answers if they know whether they will receive the program.

The timing of the follow-up survey should take into account the program needs and program effects. If a follow-up survey is conducted too soon, no effect will be found; while if it is done too late, the program may not benefit from the knowledge gained.

Existing Versus New Data

It is not always necessary to collect new data. In some cases, the data required for an evaluation already exist (box 7.2 offers suggestions for where to find it). Two types of data commonly exist and should be explored before deciding to collect new data.

In one program implemented in partnership with the local government, an NGO in Latin America experienced various delays with participant selection. Because a lot of time had passed between the selection of youth and the start of training, youth began to lose interest and drop out of the treatment group. As a result, the treatment group fell below the suitable number. In such a case the impact would have to be very large in order for it to be measureable.

BOX 7.2 Potential sources of data

Administrative data. Administrative data are usually collected by an implementing program for monitoring purposes.

Household survey data. National household surveys are periodically conducted in many developing countries. These include multi-topic surveys, such as the Living Standards Measurement Survey and the Demographic and Health Survey, which can cover a wide range of information on housing characteristics, household consumption and wealth, individual employment, education, and health indicators. Other surveys, such as labor force surveys, are more restricted in scope and sometimes cover only urban areas.

Where to look:

- Statistical institutes in the respective country
- International Household Survey Network (www.ihsn.org)
- Demographic and Health Surveys (http://www.measuredhs.com/)
- Living Standards Measurement Surveys (http://iresearch.worldbank.org/lsms/lsmssurveyFinder.htm)

Census data. Most countries conduct a population and housing census every ten years, and many conduct additional surveys. The advantage of census data is that they cover the entire population, so there are data for virtually every potential treatment and comparison observation. The drawback of census data is that it is infrequent and typically contains only a limited number of indicators, limiting their value for an impact evaluation.

Where to look: International Household Survey Network (www.ihsn.org)

Facility survey data. Facility surveys collect data at the level of service provision, such as at a school or vocational training center. National ministries, state entities, or even local authorities may compile the information. In many cases, facility-level surveys will provide control variables (such as teacher–student ratio), while others may capture outcomes of interest, such as attendance rates.

Where to look: Relevant national ministries and local representatives.

Specialized survey data. A specialized survey is one that is collected for a specific purpose, often for research on a particular topic. Many take modules from the existing national household survey and add questions on topics of interest. Coverage of specialized surveys can be quite limited, sometimes resulting in little or no overlap with program areas. Nevertheless, if the evaluation team can find existing data from a specialized survey on a topic related to the evaluation, these datasets can provide a rich collection of relevant indicators.

Where to look: Local officials, donors, and NGOs in the area of interest.

Source: Reproduced from World Bank (2007a, pp. 8–11).

First, **the necessary data may already be collected in the form of administrative and M&E data.** Depending on the questions the program wants to answer, answers may already have been collected. For example, many livelihood programs already ask information on income and employment at the start of the program, thus minimizing the need for a baseline. This information is normally only collected for those in the program, however. Data must also be collected on individuals in the comparison group. To avoid inadvertently introducing biases through inconsistent data collection, it is important that any system designed for data collection is as consistent

and objective as possible for both the treatment and comparison groups. This is often difficult to do through purely administrative data collection. Unless such a system is naturally a part of the program, it is best to use a dedicated team to collect new data on both the treatment and comparison groups.

Second, **the local bureau of statistics may have already collected data** on many of the program participants and comparison groups. For smaller programs, it is unlikely that enough people in the program have been part of an existing survey. For larger programs, though, it is likely at least some have been. It is also important to understand what data was collected and how that collection was done. Ensure that the questions asked pertain to the program that we have in mind and that they sample size was large enough to warrant drawing conclusions. Check with the local statistics bureau to confirm that the data exist and can be used.

If using existing information is not sufficient, new data will have to be collected.

Internal Versus External Data Collection Team

The collection of data is the most expensive part of an evaluation for good reason. The collection of high-quality data that can be easily analyzed is key to a successful evaluation. Without high-quality data, all of the work put into designing the evaluation may go to waste. When deciding between hiring a survey firm or collecting data with internal staff, the program must choose the method that fits its budget and ensures quality and systematic data collection. Some programs want to conduct data collection on their own since it can save money. This may work well for short, simple surveys, but it has some important drawbacks, especially for extensive data collections. Due to the complexity of collecting data and ensuring the proper logistics, it is normally not advisable to collect data with program staff. While hiring a survey firm is typically more expensive than doing the data collection internally, it means the data can be collected faster and with less work from the program office. It also ensures there is a qualified team doing the data collection. (Additional guidance on quality assurance is included under the sections Training the Fieldwork Team and Supervising the Data Collection, below.) Moreover, hiring an outside firm ensures neutrality and increases the credibility of the evaluation results.

[Online Resource]

Protocol for hiring a survey firm

http://www.iyfnet.org/gpye-m&e-resource7

[Tip]

In some cases, programs attempt to have partner implementing organizations collect data through their program staff. It is not advisable to have people who are dependent on funding conduct the data collection because there is a greater chance that the results will be biased in favor of the program. If it is decided that data collection will be done internally, it is best to do it with a separate team that is focused only on data collection and is not associated with the program.

Data Collection Process and Technique

Generally, surveys should be administered by trained personnel; self-administered questionnaires should be used only in certain circumstances. When individuals fill out surveys on their own, they often interpret questions differently from what was intended by the survey team. Trained interviewers ensure greater consistency of interpretation. Also, in many contexts, participants are not as literate as we might expect or hope, so they may require guided interviews.

There are several ways to collect and record survey responses. Paper surveys are traditional. If available, interviewers can also use cell phones (to which surveying software can be downloaded), computers, or personal digital assistants. It may also be possible to tape interviewee responses. Although technology-based tools may require some initial training (usually relatively minor), they can reduce the time needed for each interview, cut the time needed for data entry, and minimize data errors that arise from traditional data entry and processing. They can therefore save time and money, especially in larger surveys. However, one also needs to consider the appropriateness of using sometimes-expensive equipment in poor households and neighborhoods.

[Online Resource]

ICT-based data collection tools

http://www.iyfnet.org/gpye-m&e-resource2

Develop and Pilot a Survey Instrument

If the evaluation plan calls for collecting new data, it is important to choose the right data collection tool. In most cases, some sort of survey will be used, often in combination with other qualitative methods, such as focus groups or key informant interviews.

Because the survey will be the basis for collecting data about participants and the comparison group, the survey design is crucial. Although designing questionnaires may seem trivial, coming up with a high-quality survey that yields reliable results is a science and an art. Surveying adolescents and youth poses additional challenges compared with surveying adults, so it may be wise to seek support from an expert consultant for this step (see box 7.3).

BOX 7.3 Factors affecting data reliability when surveying youth

Any evaluation depends on reliable information. While research indicates that young people are generally reliable respondents, there are a number of reasons why youth may be more likely than adults to misreport or even falsify answer questions:

- **Comprehension.** Young people may have less education and relatively limited cognitive ability. Does the respondent understand the question? Is the question asked using age-appropriate language? Some questions are subtle and may be difficult for youth to understand even when asked in a simple and straightforward manner.
- **Recall.** How likely is it that the respondent remembers the events or information? This has partly to do with the reference period: how long ago the event occurred or how frequently the event occurs. In general, shorter recall periods are more accurate than longer ones.
- **Confidentiality.** Does the respondent have any reason to fear reprisal or other consequences arising from the answers he or she gives? Is the interview really being conducted in private? The interviewer must be able to convince the respondent that the information is confidential.
- **Social desirability.** Does the respondent believe that the interviewer is expecting one response or another? Can one answer be perceived as "correct?" This affects especially behaviors that are illegal, stigmatized, or subject to moral strictures. Brener, Billy, and Grady (2003) report studies showing that adolescents are more likely to report recent alcohol consumption in self-administered questionnaires than in interviews, whereas there is no difference in the responses of adults. In addition, numerous studies confirm that young people are more likely than adults to provide inconsistent answers in surveys repeated over time.
- **Exhaustion.** Although surveys among adults can take many hours to complete, young people are more likely to lose patience with long interviews. For example, the NGO Save the Children created the Youth Livelihoods Development Index, which comprises three self-administered surveys for young people ages 11–24 to elicit information about assets and competencies. The pilot test found that youth "got bored with the long questionnaire and fabricated answers" (Bertrand el al., p. 5).

Selection of sample survey instruments, including the NUSAF baseline and endline questionnaire.

http://www.iyfnet.org/gpye-m&e-resource11

Note: The NUSAF questionnaire is very long and, although it was based on a previous survey, took one full-time worker four weeks to pretest. Although most surveys will not contain so many questions, it offers a good example of the types of questions that can be used in youth livelihood programs. It is also important to recognize that many outcomes may not be easy to measure (e.g., risky behaviors, mental health, empowerment). Different surveys use different approaches, and it is recommended to use previously developed instruments—ideally surveys that are scientifically validated—for guidance.

Designing and Testing the Survey

Before the survey can begin in the field, the questionnaire must be developed. This is done through an iterative process that will usually take one to two months.

Step 1: Design

The questionnaire is based on the outcomes and indicators previously developed.

Local language, dialects, and youth slang are important aspects to incorporate, and a translator may be needed to do this well. If sensitive topics are included in the questionnaire, such as questions about mental health or violence, questions must be formulated thoughtfully and in line with local norms and customs. The first draft will usually contain questions that will eventually be cut or changed.

Step 2: Internal Review

Once a questionnaire has been drafted, other team members and stakeholders such as the program manager, M&E officer, principal investigator, and fieldwork team should review it to confirm that the questionnaire collects all the information needed.

Step 3: Piloting

The draft questionnaire is then taken to the field. The importance of this step is often overlooked, but it is critical for the production of a quality evaluation. Field-testing is crucial to confirm that the survey's length, formatting, and phrasing are all appropriate, and to make sure that the survey can yield consistent and reliable results. The questionnaire should be tested on a selection of individuals who are similar to those who will be in the program, but who will not be in the final sample. This will ensure that those people who receive the final questionnaire are not influenced by having already been exposed to the questions. It is also important to pretest the procedures that will be used for locating interviewees to ensure that they can easily be found.

Step 4: Revision

The draft questionnaire is revised to address the issues raised in the field. If necessary, the steps can be repeated until all issues have been resolved.

Training the Fieldwork Team

When the questionnaire is ready, the fieldwork team must be trained to administer it. The survey expert or data collection firm should develop a manual to be used as a training tool and reference guide for interviewers. At a minimum, the manual should discuss the survey objectives and procedures, including procedures for dealing with difficulties in the field. Each survey question should be explained so that interviewers understand the rationale for the question's inclusion in the survey. In addition, the manual should provide interviewers with specific instructions on how to ask each question and obtain usable information. The principal investigator and program manager should review the manual. Box 7.4 presents a sample outline of a survey manual.

[Tip]

Good practices for surveying youth include the following:

- Obtain informed consent from both the young person and the parent (see section below on human subjects protection).
- Use familiar local language or slang, if appropriate.
- Be mindful of the young person's attention span; keep surveys short and interesting.
- Use probing questions to improve the quality of responses; refer to the recent past to help with memory and recall.
- As with all respondents, be cautious about the timing and phrasing of sensitive questions.
- To help with finding youth later, gather a lot of information on family, friends, and neighborhood contacts.
- If information about the household is needed, include a separate survey module targeted at parents or guardians.

[Online Resource]

Training manuals for data collection

http://www.iyfnet.org/gpye-m&e-resource12

BOX 7.4 Sample outline of a survey manual

1. Objectives of the survey
2. Duties, roles, and expectations of interviewers, supervisors, and other survey personnel
3. Procedures for checking data accuracy
4. Detailed survey and interview procedures (including procedures for identifying, locating, and contacting respondents, as well as information on surveyor conduct, confidentiality, objectivity, interview pace, bias, and probing)
5. General instructions for filling out the questionnaire and coding
6. Simple explanations of each question
7. Instructions for finishing and checking the survey and thanking respondents
8. Instructions for filling out the field report and notifying supervisors of any difficulties encountered

[Tip]

Be mindful of cultural norms and local customs when recruiting and assigning interviewers. For example, it is usually a good idea to use female enumerators to interview female respondents, particularly when sensitive questions are being asked. If respondents (or their guardians) do not feel comfortable with an enumerator, it is more likely that they will not participate in the survey, or, if they do, that the information provided will be incomplete, inaccurate, and therefore unreliable.

Training interviewers can take a few days or more than a week, depending on the complexity of the survey. Training should begin by going through the entire survey, question by question. Then, each interviewer should practice on another interviewer. Interviewers should be encouraged to ask questions during this process to ensure everyone understands each of the questions. This process should continue until all interviewers are very familiar with all questions. After the training is complete, interviewers should be taken to a site where they can practice the questionnaire on at least five people who resemble the sample respondents.

Interviewer training is both a training process and a job interview. Invite at least 20 percent more interviewers to the training than are expected to be needed, and accept only the best.

If a survey firm is contracted, they will be in charge of the training. It is often a good idea to have someone from the program attend the first few days of the training to answer questions that arise. This is the last chance to eliminate errors in the questionnaire.

Human Subjects Protection

Research that involves human beings can sometimes create a dilemma. When our research is intended to generate new knowledge for the benefit of a specific program or an entire field, for example by measuring the impact of a youth livelihood intervention, we may be inclined to consider the outcomes of our evaluations to be more important than protecting individual research participants. Clearly, we should not use young people solely as means to an end, and there are procedures in place to help us assess our evaluation's ability to protect participants.

Basically, three main principles protect the interests of research participants (NIH 2008, pp. 17–20):

- **Respect for persons.** This principle refers to making sure that potential participants comprehend the potential risks and benefits of participating in the evaluation. In practice, this means that a process must be in place to ensure informed consent, the explicit willingness of young research participants to answer the survey questions in

light of their clear understanding of the nature of the survey.

- **Beneficence.** This principle refers to doing no harm and maximizing the possible benefits of the research.
- **Justice.** The principle requires that individuals and groups be treated fairly and equitably in terms of bearing the burdens and receiving the benefits of research.

In order to ensure the highest ethical standards in an evaluation, many researchers will be required to submit their impact evaluation plan for a review by an institutional review board (IRB) in the donor country, the host country, or both. These reviews are mandated by law for anyone engaging in research supported by the U.S. government and many other governments as well as most universities throughout the world. Even if they are not legally required, conducting ethics reviews is a good idea for anyone working with human participants. Ideally, the IRB would review the survey before it is piloted, but certainly before the final survey is implemented at large. IRBs can be found in any U.S.–based university (the best option when working with a U.S.–based researcher) or through a local ethics review board. Other institutions, such as the U.S. National Institutes of Health or Innovations for Poverty Action also conduct ethics reviews on request. Box 7.5 shows a sample outline of an IRB application, and box 7.6 provides advice on the IRB approval process.

[Definition]

An **institutional review board**, also known as an independent ethics committee, is a committee that has been formally designated to approve, monitor, and review research involving human participants with the aim to protect the rights and well-being of these individuals.

Informed consent refers to the explicit willingness, preferably in writing, of a person (and, when necessary, his or her parent or guardian) to participate in the research. Informed consent requires full information about all features of the research that may affect a young person's willingness to participate.

BOX 7.5 Sample IRB application format

Title of Study: _____

Country and Location: _____

Anticipated Start Date and End Date: _____

Investigator(s), including name, position, department, and institution of each: _____

I. Purpose/Background/Significance of the study, including why it is valuable.

II. Study design, including how treatment and comparison groups are determined and timing of the program. Describe all measures to be collected.

III. Describe study participants and if any are a vulnerable population. Note if there is to be any compensation to participants.

IV. Describe informed consent process.

V. Are there any possible risks or benefits of the study?

VI. How will confidentiality be maintained?

VII. Misc.: Memorandum of Understanding or letter of support from partner organization(s), survey(s), consent form(s), certificate of human subjects training (NIH or equivalent) for all research personnel.

Source: Adapted from Innovations for Poverty Action (2010).

BOX 7.6 Advice on the IRB approval process

When your organization has no approved IRB

Almost all academic institutions have IRBs, as do a number of donor agencies and international NGOs. If you are working in partnership with one of these agencies, you may be required or encouraged to follow their procedures for obtaining IRB approval. If you are working independently or have no access to a partner's IRB, many universities and other institutions provide ethics review services. The Office for Human Research Protections of the U.S. Department of Health and Human Services maintains a searchable database of more than 8,000 IRBs around the world, from Afghanistan to Zimbabwe (see http://ohrp.cit.nih.gov/search/irbsearch.aspx?styp=bsc.) In addition, many independent agencies provide ethics reviews, generally for a fee. For more information, see the Association for the Accreditation of Human Research Protection Programs (http://www.aahrpp.org/www.aspx), and the Consortium of Independent Review Boards (http://www.consortiumofirb.org/).

When there is not enough time to go through a full IRB approval process

First, reassess the probability of obtaining a review in the time available. Your program is intervening in the lives of young people and their families, and you have a responsibility to ensure that your participants are protected, as well as you possibly can, from harm. However, IRB approvals can take up to several months, and you may be rushed to begin implementation. If, after careful analysis, there is indeed no possibility of obtaining timely IRB clearance, at minimum all members of the evaluation team should have been trained on the protection of human participants in programs and research. The National Institutes of Health (NIH) offers free online training (in English and Spanish). For more information, see: http://grants.nih.gov/grants/policy/hs/index.htm

While respecting ethical standards is essential in all research projects and evaluations, special issues may arise when working with young people that require additional attention (see table 7.3). These issues make the involvement of an IRB even more critical than in other evaluations, and require that the researchers and consultants engaged in the evaluation receive explicit training on child and youth development prior to beginning the evaluation. In addition, clear protocols should be developed to define what information will be collected and how it will be used in order to maintain the highest ethical standards and protections for the participants. For an example of applying human subjects protection standards in Honduras, see box 7.7.

TABLE 7.3 Overview of ethical considerations when conducting research on children and youth

Issues	Why it Matters	What to Do
Information about Risks and Benefits of Participation	Young people may have a different ability than adults to accurately assess the benefits and risks associated with participating in a particular program or research initiative. They may also be more risk-taking in general, making them more vulnerable to the potential negative consequences of participation.	• Anticipate possible consequences for the children and youth involved. Do not proceed unless potentially harmful consequences can be prevented or mitigated. • Provide young participants with an explanation of the proposed research objective and procedures in a language and format appropriate to their age, maturity, experience, and condition. • Provide explicit discussion of any inconveniences or risks the young person may experience if she or he agrees to take part in the program or evaluation. • State clearly that there is no obligation to participate in the study and that the decision to participate in the study will have no effect on eligibility for the program. • Do not raise unrealistic expectations about the benefits or rewards to participation. • If any, provide only modest rewards or incentives to participate that are in line with local living standards.
Consent	Young people may not have reached the age of legal maturity; their parents or guardians need to be asked for consent prior to engaging the youth themselves. Moreover, obtaining young people's truthful opinion can be difficult because they are often socialized into complying with adult opinions, regardless of whether or not they agree.	• Determine the age of majority in the country and consult locally to determine who must give permission to work with the young people (parents, teachers, local authorities, community leaders, etc.). • When working with minors, always seek informed consent from parents or guardians. • If age, maturity, and situation of the young participants allow, also obtain informed consent from the youth in addition to that of their parents.
Data Collection	The collection of information on sensitive topics (e.g., drug use, sexual activity, involvement in crime) or distressing experiences (abuse, loss of parents, deprivation) is more delicate when dealing with children and youth compared to adults. Their emotional and physical vulnerabilities have to be protected.	• Prior to interviewing young people, try to collect as much information as possible from alternative indirect sources (adults, administrative records, etc.). • Consult locally and design questionnaires, focus group guidelines, and other materials according to the characteristics of the specific target group (e.g., make sure that survey instruments are age-appropriate and comprehensible). • When necessary, acknowledge that questions can be sensitive, and anticipate and address the concerns of parents and participants. • State clearly that the young participant can refuse to answer any or all questions, and that this will have no effect on eligibility for the program. Such disclaimers should be repeated before asking sensitive questions.
Confidentiality and Protection	Protection of privacy is always crucial, and even more so when dealing with young respondents and sensitive topics. Given the involvement of parents or other guardians during the consent process and as legal representatives, there may be tradeoffs between confidentiality and the ethical obligation to protect the safety of the respondents that do arise when working with adults. For example, the presence of parents in the interview may undermine the privacy of the youth. At the same time, there may be a responsibility to inform guardians if the young person is at risk of harm.	• Always ensure the privacy and confidentiality of responses from parents and young participants, which will also strengthen the reliability of the information provided. • Never release information about the respondent without the express approval of the respondent and his or her parent. • Plan how to intervene if the respondent provides information suggesting they or others may be at risk of harm (from domestic abuse, neglect, crime and violence), or may require medical, legal, or other services. • At the beginning of each interview, and regardless of the apparent conditions of the respondent, inform *all* participants of the resources available for referral.

BOX 7.7 Human subjects protection in practice

To conduct a survey for the job-training program *Mi Primer Empleo* targeted at urban youth in Honduras, the World Bank contracted the National Opinion Research Center (NORC) at the University of Chicago to adapt questionnaire design and manage the data collection process. Even though Honduras does not have any statutory requirements for dealing with sensitive survey data involving human participants, the terms of reference for the evaluation required U.S. IRB approval for the research design and data collection plan, as well as data security procedures that meet international standards. NORC therefore submitted all research protocols and questionnaires to its university IRB for approval prior to beginning fieldwork.

Given the nature of the research, field interviewers and supervisors were screened regarding their experience with youth-related surveys. During the program registration process, applicants were informed that they would be asked to participate in a voluntary survey but that their decision to participate in the survey would in no way influenced their selection for the training programs. Given that the legal age of consent is 18 years in Honduras, the data collection team sought written consent from respondents aged 17 or younger, and oral or written consent from the minor's parent or guardian for program registration, as well as a separate consent from the minor and the guardian to participate in the evaluation survey.

To ensure confidentiality, personal information was strictly separated from interview forms, and the latter contained only a numeric identifier. Thus, personal registration information (names, address, etc.) was available exclusively to the implementing organization (Ministry of Labor and Social Security) for the purpose of contacting youth who had registered, while response data (without personal information) was delivered only to the World Bank for analysis.

Source: NORC (2007).

Conduct Baseline Survey and Analysis

The baseline survey is the first data collected on the treatment and comparison groups. As discussed previously, a baseline is not always necessary for all programs and impact evaluation methods. However, collecting baseline data is highly desirable because it provides an early assessment about whether the chosen impact evaluation design is valid in practice, while providing useful information about beneficiary characteristics that can inform the program.

Another good reason for conducting a baseline survey is that it may help locate participants later on. The baseline survey, if conducted, should always include a list of contact information from the person surveyed, and also from friends and family who can be called during the follow-up survey.

[Tip]

For detailed guidance on ethical approaches to research involving children and youth, consult

Society for Research in Child Development. 2007. *Ethical Standards for Research with Children*. Available at http://www.srcd.org/index.php?option=com_content&task=view&id=68

Schenk, K. and Williamson, J. 2005. *Ethical Approaches to Gathering Information from Children and Adolescents in International Settings: Guidelines and Resources*. Washington, DC: Population Council. Available at http://www.popcouncil.org/pdfs/horizons/childrenethics.pdf

Timing

Baseline data should be collected shortly before the program begins. If it were to be conducted after program initiation, the program may have already influenced characteristics measured. If the baseline survey were conducted much in advance of the program, the information collected may not accurately reflect the situation of participants at the beginning of the intervention.

If we are doing a prospective evaluation, individuals will need to be assigned to treatment and comparison group before the program begins. However, that assignment decision should not be communicated to the survey participants until after the baseline data has been collected.

Supervising the Data Collection

Quality assurance is key to ensuring that the data collected is of the highest quality. First, it is important to conduct validity testing to ensure interviewers are meeting the standards of their job and that they meet the target number of surveys per day. It is customary to establish an independent team to audit 10–15 percent of the surveys to verify that respondents exist and that data was collected accurately. Incentives may help ensure that interviewers keep a positive attitude in a difficult job. In addition to wages, interviewers often receive a per diem allowance to cover food and housing while traveling, as well as other incentives.

Second, steps should be taken to protect the data collected. Information can be lost if completed questionnaires are misplaced or computers are stolen or malfunction. To avoid the loss of data, surveys should be collected as soon as possible from interviewers and stored safely. Computer data should always be backed up.

Finally, it is important to ensure quality data entry. Using electronic data entry tools such as cell phones or personal digital assistants can help avoid data entry errors, as can standard quality control measures, such as entering the same data twice.

Analysis and Report

Once the baseline data has been collected, the lead evaluation expert and the research assistant should complete the baseline analysis and report. As there are not yet program results to report, the baseline report will consist of descriptive statistics. The average values of the demographics of treatment and comparison groups should be compared to ensure the similarities between the two groups, and statistically significant differences should be noted. Any issues that arose with data collection should also be presented in the baseline report (see box 7.8 for a sample outline).

BOX 7.8 Outline of a baseline report

1. Introduction
   - 1.1 Description of Program and Evaluation
   - 1.2 The Research Team
   - 1.3 Report Overview
2. Background
   - 2.1 Setting and Location
   - 2.2 Historical Background
   - 2.3 Scientific Background
   - 2.4 Program Description and Implementing Partners
3. Intervention
   - 3.1 Group and Participant Selection
   - 3.2 Description of Intervention
   - 3.3 Issues with Implementation
4. Impact Evaluation Design
   - 4.1 Intervention Objectives and Hypothesized Outcomes
   - 4.2 Research Design and Randomization
   - 4.3 Outcome Measures
     - 4.3.1 Primary Desired Outcomes
     - 4.3.2 Secondary Desired Outcomes
     - 4.3.3 Adverse Outcomes
     - 4.3.4 Other Measures of Interest
     - 4.3.5 Treatment Heterogeneities
   - 4.4 Problems Encountered
   - 4.5 Intervention and Evaluation Flow Chart and Timeline
5. Baseline Survey Administration
   - 5.1 Individual and Group Surveys
     - 5.1.1 Baseline Survey Development and Pre-testing
     - 5.1.2 Enumerator/Survey firm Recruitment and Training
     - 5.1.3 Baseline Survey Implementation
     - 5.1.4 Problems and Concerns
   - 5.2 Other surveys
6. Baseline Analysis
   - 6.1 Baseline Characteristics of Participants
   - 6.2 Power Calculations and Tests of Balance on Baseline Data
   - 6.3 External Validity
   - 6.4 Data Quality Issues
7. Conclusions
   - 7.1 Discussions
   - 7.2 Interpretation
   - 7.3 Generalizability

Appendix

Source: Based on Bose (2010).

As we have seen in note 6, the validity of each impact evaluation method rests on a number of assumptions. The baseline analysis can play an important role in verifying these assumptions to confirm that our evaluation method of choice can be used, or, if problems are encountered, how to resolve the issue. Appendix 3 provides a list of verification and falsification tests that can be used to assess whether the assumptions underlying our desired evaluation hold true.

Conduct Follow-up Survey and Analysis

When an evaluation method rests on collecting new data, the follow-up or endline survey will provide the long-awaited data that will allow us to analyze whether our intervention was successful or not. When an evaluation is based fully on existing data, then its analysis will be conducted during this stage.

Timing

The program manager and lead evaluator will jointly determine the timing of the follow-up survey. Not every program benefit will be observable immediately after the intervention, so the follow-up survey must be conducted after enough time has passed for the impact to materialize. The time varies according to program and depends very much on the specific outcomes of interest. For example, young people participating in a training program may actually face a short-term disadvantage in terms of earnings compared with their peers, since they cannot work during the time they are in class. However, if our training provides relevant skills, we would expect them to have a relatively higher income over the medium- to long-term. The timing of the follow up will be crucial to identifying the true effect of the intervention.

If we want to measure both short- and long-term outcomes, we may need to conduct several follow-up surveys. Although this will increase the cost of the evaluation, it may also drastically enhance its value. Impact evaluations that follow treatment and comparison groups over many years are relatively rare, and their results are all the more demanded and appreciated. Conducting more than one follow-up survey will also allow us to analyze how the program outcomes change over time. However, if program implementation is delayed, we may be left with too little time between the end of the program and the end of our budget or grant cycle to conduct a follow-up survey that will capture long-term outcomes. It is therefore important to realign the evaluation cycle if changes in the implementation timeframe occur.

Tracking

One major difference between the baseline and endline surveys is the issue of tracking respondents. If the surveyed youth are not found at follow up, it can introduce very serious biases to the analysis and reduce the value of findings. For instance, if participants who perform the worst drop out, the evaluation results will likely overestimate the impact of the program. But it may also be that the most able youth drop out. Because we don't know for sure whether attrition will lead us to underestimate or overestimate impact, minimizing attrition is essential to conducting any good evaluation. Although it is almost never possible to find 100 percent of individuals previously surveyed, every effort must be made to find as many as possible. A generally acceptable rate of attrition is 5–15 percent, meaning that at least 85 percent of youth in both the treatment and comparison group should be located.

[Tip]

To ensure that final evaluation results are considered reliable later on, it is good practice to include external experts in the review process for the baseline and final report. Moreover, by disseminating the baseline report, program and evaluation staff can create public interest in the ongoing research and strengthen the ownership and dialogue among internal and external stakeholders.

[Tip]

It is often possible to identify intermediate indicators that are consistent with the anticipated long-term outcomes. For example, the impact of entrepreneurship education and promotion programs on the probability of starting a business might not always materialize for a number of years (students leave school, get a job to gain relevant experience, and eventually consider starting their own business.) By measuring short- and medium-term outcome indicators, such as business skills, the preference for starting a business as a career choice, and concrete steps taken toward starting a business, it is possible to obtain intermediate impact results without having to wait several years.

[Definition]

Attrition refers to the dropout of participants or survey respondents. This represents a problem for the evaluation because the dropouts are likely to be systematically different from those who can be found, thus skewing our results. Attrition can occur for any number of reasons, such as loss of interest in the program, migration, or simply the unwillingness to participate in the survey.

In the Middle East, the Syria Trust provided mobile phone charge cards to motivate youth to participate in a survey. To save costs, Syria Trust asked mobile phone operators to provide these cards as in-kind donations. Mobile phone companies provided 10,000 cards at US$2 each, a value of US$20,000). For the phone companies, it was good publicity at minimal cost.

In Uganda, the NUSAF program hired a firm to conduct a 10-minute tracking survey of respondents one year after the baseline and one year before the endline. The questionnaire asked participants who could be easily located for their updated contact information. For those who could not be easily found, information was collected from friends and family on the likely whereabouts of the person. This information was then kept for the endline to aid the teams in finding survey respondents, as well as giving the team an indication of how hard or easy it will be to find people.

[Tip]

Additional ways to facilitate tracking include the following:

- Ask the advice and help of local leaders, officials, and residents. Locals may know the best way to find someone.
- Involve field enumerators from the study location since they are familiar with the area and local customs.
- If participants still cannot be found, select a random sample of those not found to conduct a very aggressive search for them. If selected randomly, those who will be eventually found through more intensive search can be considered representative of others who have not been found.

Tracking people, especially highly mobile youth, can be difficult. The following are three common ways to reduce attrition:

- **Gather good contact information during baseline.** The baseline survey should include various types of contact information (street address, email address, phone number, etc.) from the respondent and also from friends and family who can help locate the youth for the follow-up survey. Using social media channels such as Facebook can also help to keep track of young people.
- **Motivate youth in treatment and comparison groups to be available for future surveys.** Incentives to participate in follow up can include small payments for their time or lotteries for cash or prizes. Youth can be notified of these incentives through prearranged communication (perhaps at baseline), or through mass media, such as radio and newspaper advertisements.
- **Use a tracking survey.** For evaluations that have a lot of time between the baseline and endline, such as two years or more, and especially for those that do not use a baseline, a short, very fast tracking survey can be used to estimate the likely attrition and gather additional information. If the program is budget-constrained, one may also consider doing follow-up surveys by telephone to get up-to-date contact information from survey respondents, while limiting personal visits to those youth who cannot be reached over the phone.

Follow-Up Survey Design and Data Collection

It is likely that the program or evaluation team will want to add a few additional questions to the original survey (see box 7.9). These may include questions about attendance, dropout, and motivations for both, since this information can be used to estimate how much treatment individuals actually received. New questions will need to be piloted and revised as necessary. In general, it is best to keep follow-up questions and the order of questions as similar to the baseline survey as possible to ensure they are comparable. Unless there was a major issue with a question in the baseline survey, it is best to leave it worded the same in follow-up surveys. The survey manual will also need to be updated to reflect any changes from the baseline. In particular, it should include specific protocols for tracking survey participants.

BOX 7.9 Common types of questions to be added to the follow-up survey

- Reasons for not participating or dropping out
- Frequency of participant attendance or amount of benefits received
- Participant satisfaction with the program
- Participant rating of quality of program
- Participant self-assessed outcomes of the program

Finally, interviewers will need the same level of training and oversight as with the baseline survey to ensure the best quality of data collection. If possible, select the best interviewers from the baseline staff to conduct the follow-up survey. Interviewers with high error rates or those who were less reliable should be replaced or given additional training.

Final Analysis and Evaluation Report

After follow-up data is collected, the final impact evaluation report can be produced, which represents the main product of the evaluation. The final report will repeat much of the information presented from the baseline survey, and it will add detailed information on the endline survey administration and final data analysis.

The analysis will be based on the outcomes and variables previously identified. In some rare cases, the analysis can be done by a simple comparison of the average values between the treatment and comparison groups (usually in the case of lottery designs). In practice, however, one will often use some form of *regression analysis* to control for several key variables that may otherwise bias the results.

Box 7.10 presents a sample outline for sections of an evaluation report that can be added to the baseline analysis. All of this information is important to ensure that someone not involved in the evaluation can interpret the results correctly.

BOX 7.10 Example of additions to baseline report after endline

7. Endline Survey Administration
   - 7.1 Endline Individual and Group Survey
     - 7.1.1 Endline Survey Development and Pre-testing
     - 7.1.2 Survey Firm/Interviewer Recruitment and Training
     - 7.1.3 Mobilization and Tracking Protocols
     - 7.1.4 Endline Survey Implementation
   - 7.2 Qualitative Protocols
   - 7.3 Problems and Delays
   - 7.4 Data Quality Issues
8. Data Analysis
   - 8.1 Statistical Methods Used
   - 8.2 Levels of Analysis
   - 8.3 Summary of Outcomes
   - 8.4 Ancillary Analyses
9. Conclusions
   - 9.1 Discussions
   - 9.2 Interpretation
   - 9.3 Generalizability
   - 9.4 Directions for Future Research

Appendix

Source: Based on Bose (2010).

Understanding Heterogeneity

Not all program beneficiaries may benefit from our intervention in the same way. Therefore, one important value of evaluation is to understand the variation in program impacts. For instance, many programs want to know whether boys or girls, younger or older youth, or those with higher or lower levels of education or experience perform better in the program. In addition to looking at gender, age, or education, we may also want to assess whether outcomes differed by participants' initial wealth (the value of

[Definition]

In statistics, **regression analysis** includes any techniques for modeling and analyzing several variables. In impact evaluation, regression analysis helps us understand how the typical value of the outcome indicator changes when the assignment to treatment or comparison group is varied while the characteristics of the beneficiaries are held constant.

[Online Resource]

Impact evaluation reports

http://www.iyfnet.org/gpye-m&e-resource13

[Definition]

Impact heterogeneity refers to differences in impact by type of beneficiary; that is, how different subgroups benefit from an intervention to a different extent.

participant assets), social capital (access to networks), or psychological traits (optimism, risk attitudes, and the like). Understanding which participants have benefited the most and which the least from our program can help us better design or target the intervention. (For more information on measuring heterogeneity, see Measuring a Variety of Impacts in note 8.)

For example, if our evaluation finds that a livelihood training program had a greater impact on men, future iterations of the program could focus more on men to increase the overall return of the program. Alternatively, depending on priorities, we could explore ways to get women more involved so that they, too, benefit from the program.

However, as is noted in box 7.11, heterogeneities of interest should be specified in advance of any analysis and all results should be reported, not just those found to be statistically significant. We want to avoid data mining, which can be an especially big problem with heterogeneity analysis.

Bruhn and Zia (2011) studied the impact of a comprehensive business and financial literacy program on firm outcomes of young entrepreneurs in an emerging postconflict economy, Bosnia and Herzegovina. Although they did not find significant average treatment effects of the training program on business performance, they identified high levels of heterogeneity among participants. Specifically, young entrepreneurs with relatively high financial literacy prior to the program were found to exhibit improvements in sales due to the training program. The effects on profits were also positive for this sub-group. The results suggest that training should not be the sole intervention to support young entrepreneurs and that the content of the specific course may have been appropriate for a very specific set of young entrepreneurs, but not for all.

BOX 7.11 Data mining

Data mining is a serious problem within statistics. It is especially problematic with very long surveys that ask many questions, often in different ways.

In data mining, a person seeks out results that confirm specific beliefs about a program and ignores results that do not confirm these beliefs. For instance, a program officer may strongly believe that a training program has a positive impact on youth. Once the officer receives the data from the evaluation, she finds that there is a statistically significant increase in time spent working, but the youths' average income is not statistically higher. Reporting only the increase in time spent working and not the fact that there is no change in income is a kind of data mining.

Data mining can happen in two ways. The first is when we ignore evidence that is counter to our beliefs and report only those that confirm our beliefs. The second is a statistical anomaly. In statistics, there is always a chance that a variable will be found significant. In fact, at least 5 percent of the time, something will be found to be significant that is in fact not significant. If an evaluator collects 100 pieces of information, at least five will be incorrectly attributed to be significant, when they are not. If the researcher looks for these five, and reports only these five, then the results are, in fact, incorrect.

An evaluation may find no statistically significant impact from a program. But by exploring every possible heterogeneity it is very likely that, due to statistical randomness, researchers will find some impact on a group. To avoid data mining, we should identify all of the outcomes of interest before conducting the analysis, and report all of these outcomes without fail, including those where no impact was found. In this way, the whole picture can be understood.

Interpretation of Results

Quality of implementation: Results depend a great deal on how well an intervention was implemented. The final evaluation report should therefore discuss the quality of the implementation in detail. Having good knowledge of how the program was implemented is particularly important when evaluation results show a limited or negative impact since it allows us to differentiate problems with implementation from problems with program design. In order to be able to accurately interpret the evaluation results, it is necessary to embed the impact evaluation in a framework of strong monitoring, process evaluation, and other qualitative tools.

Generalizability of findings: Ideally, our impact evaluation has external validity, which means we can generalize our findings to other populations and conditions. Whether this is the case largely depends on the sampling strategy chosen in the evaluation. The more representative the sample, the more confident we can be that a program would also work with different or larger groups of beneficiaries. This has important implications in terms of scalability and replication of the intervention. In general, it is prudent to assume that changes over time, different environments, and different delivery mechanisms from one site to another have the potential to significantly affect the impact of the program in either direction. We should therefore always be careful when translating evaluation lessons from one program to another and be mindful that monitoring and evaluation will always be necessary for continuous learning and program improvement.

Disseminating Findings

Once the results of the impact evaluation have been obtained, the final step is to disseminate the results to program staff as well as to those outside the program who may be interested in the results.

Internal Dissemination

Internal dissemination of an evaluation provides the basis for organizational learning. Sharing results with the program staff and the rest of the organization fulfills one of the main motivations for conducting an evaluation in the first place: enhanced program management (see note 1). In order to generate interest and ownership, the process of internal dissemination is best started immediately after the baseline is completed—for example, by sharing and presenting baseline findings. The results of the evaluation should then be disseminated to executives and leaders in country offices and headquarters, where applicable. The report could include a discussion about how the results can affect the design of future or current initiatives.

External Dissemination

Dissemination should also target external stakeholders, such as local authorities, national ministries, local and international NGOs, universities (especially the development, economics, and public health departments), multilateral organizations (such as the UN, World Bank, and regional development banks) or bilateral donors (e.g., USAID, GIZ, DFID). Indeed, impact evaluation findings are generally in high demand, especially in the youth livelihood field, where rigorous evidence on what works and what doesn't is still scarce. There are numerous ways to reach external audiences, and dissemination plans typically use online and face-to-face channels (see box 7.12). Evaluation findings that are shared widely can have ripple effect throughout the world.

[Tip]

Having good attendance data from program monitoring is extremely useful as it tells us not only how many youth were enrolled but also the extent to which the services offered were used. This allows distinguishing between regular and irregular participants and identifying if someone drops out in the middle of the program (possibly replaced by someone else). If this information is not collected and analyzed, it is likely that an impact evaluation will underestimate program effectiveness. Such information also helps us understand the effect of different dosages; for example, the difference in outcomes for someone who received 100 hours of training versus someone who received only 50 hours.

BOX 7.12 Selected dissemination outlets

Online dissemination

- Organization's Web site
- Newsletters
- Online knowledge portals (to upload the report and results)
  - Youth Employment Inventory http://www.youth-employment-inventory.org/
  - Youth Employment Network Groupsite http://yenclinic.groupsite.com
  - Eldis http://www.eldis.org/
  - Zunia http://zunia.org/
- Research paper databases
  - IZA Discussion Papers http://www.iza.org/en/webcontent/publications/papers
  - Social Science Research Network http://papers.ssrn.com/sol3/DisplayAbstractSearch.cfm
  - EconPapers http://econpapers.repec.org/
- Blogs and social media

Face-to-face dissemination

- Thematic conferences
  - Global Youth Economic Opportunities Conference http://www.youtheconomicopportunities.org
  - SEEP Annual Conference http://www.seepnetwork.org/Pages/conference.aspx

Presentations

- International Organizations (World Bank, IDB, OECD, ILO, UNICEF, UNDP, etc.)
- Bilateral Donors (USAID, GIZ, DFID, AfD, etc.)
- Universities (local and international)

[Online Resource]
Examples of collateral products

http://www.iyfnet.org/gpye-m&e-resource8

Collateral Products

Policy Briefs

Policy briefs help communicate the results to internal and external stakeholders. A policy brief (often no more than four pages) presents the core findings of the evaluation in a plainly written format that includes graphs and charts and that makes programmatic and policy recommendations.

Working Papers

The evaluation expert may work with the program team to write working papers and articles for publication in academic journals and to present research findings at universities and research institutions. Working papers can then be published and disseminated through the academic associations to which the investigators belong. Being cited in academic papers is a great way to increase the visibility of the program and to create interest among donors.

Troubleshooting

As with any program or evaluation, it is common to encounter problems when conducting an impact evaluation. The following list provides examples of common issues at the different steps in an impact evaluation and how to avoid or mitigate them.

Preparing for the Evaluation

Wrong program to evaluate. A lot of money can be wasted on impact evaluations whose benefit and contribution are unclear. Given limited resources, it is important to target impact evaluations at strategic and untested interventions with potential for replication and scaling up.

Unrealistic objectives. Many interventions suffer from "mission drift," whereby the expressed objective of a program changes as time goes on. It is difficult to establish useful evaluation indicators under such circumstances. Similarly, stating unrealistic objectives in terms of intended outcomes is likely to result in evaluation findings that show no impact on these outcomes. It is important to be realistic when defining the desired outcomes and learning objectives of the evaluation.

External influences. Even after agreeing to a specific evaluation design, political factors may impede moving ahead with the selected evaluation strategy. Alternatively, external factors can rush or delay implementation, affecting the delivery of services and the evaluation, such as through delayed or inconsistent treatment, or the contamination of treatment and comparison groups. One possible way to reduce the influence from third parties is to firmly agree on an implementation and evaluation plan (ideally a memorandum of understanding) and to revise it periodically.

Defining Timeline and Budget

Unrealistic planning. When developing the timeline and budget, the main risk is to underestimate the time and resources needed to carry out an impact evaluation properly. It is common to experience delays in program design and implementation, which, in turn, will also increase the duration—and probably the cost—of the evaluation. For example, delays can result in key staff and consultants being no longer available. Conservative budgeting and forward looking staffing is essential.

Setting Up an Evaluation Team

Recruitment. Recruiting a good impact evaluation team, from writing the terms of reference to identifying qualified experts and firms, can be a challenge. Underestimating the expertise needed in different stages and hiring the wrong people can lead to significant delays and cost overruns, and ultimately impair the results of the evaluation. It is necessary to ensure that the requirements for each role are clearly defined up front and fulfilled by the respective expert or firm. Working with established institutions (such as universities and think tanks) that have a track record in conducting quality research studies can help build local support and ensure that the final results are widely accepted.

Changing staff. Firms that win evaluation contacts sometimes replace key staff

with less experienced personnel. This can be prevented through clear contractual clauses with penalties or remedial actions.

Survey team management. Managing an internal survey team becomes complicated very fast. When doing data collection with program staff, make sure to understand the full staff needs and ensure there is enough oversight and management in place to handle the team.

Developing an Evaluation Plan

Limitations of existing data. When working with secondary data, it is important to ensure its availability and quality. Existing surveys may not ask the questions relevant to our particular evaluation or address our population of interest, or they may have a sample size too small to adequately power our study. Before committing to using only existing data, it is important to fully understand its limitations.

Disconnect between program and evaluation. Insufficient communication and coordination between the implementing organization and the lead evaluator can result in choosing an evaluation design that will not be feasible in practice. Keeping key program staff involved in the evaluation planning can help ensure the evaluation suits the operational context. If a disconnect does arise and it is caught in time, the best solution is to find a more realistic evaluation method.

Selection bias. Carefully identifying the sample, and randomizing study participants is the simplest and most robust way to eliminate selection bias. If selection bias is present in the data, one imperfect solution is to compare the outcomes among the treated group to a matched sample drawn from a different dataset.

Developing and Piloting a Survey Instrument

Measuremania. Targeting too many outcomes and thus including too many questions in the survey instrument often extend the cost of the survey beyond the survey budget. Too many questions increase the burden on survey participants and may reduce response rate and the quality of responses. Cutting questions related to indirect outcomes is a good way to limit this issue.

Insufficient testing. The step that often gets skipped in the interest of time is piloting the evaluation tools. Piloting is a critical step in the process that cannot be eliminated, especially because surveying youth poses additional challenges that may not be immediately understood. If the tool isn't validated, the results could be inaccurate, incomplete, or misleading. Take the time necessary during the field-testing phase of a survey to ensure that the information collected is of the highest quality.

Discounting ethics. Administering a survey that hasn't been approved by an IRB or local ethics committee may lead to massive pushback from stakeholders and may disqualify the entire evaluation. Basic ethics training for all parties involved in the evaluation is a minimum requirement.

Conducting a Baseline Survey and Analysis

Finding respondents. It may be difficult to locate youth for the survey. In this case, it is advisable to involve local program staff and other stakeholders in finding these participants.

Data quality. Even professional survey firms may not always have a good understanding of impact evaluation and may not be as qualified and reliable as one may hope. Interviewers may falsify or incorrectly record information. Poor data collection methods should not be tolerated. If contrived or low quality data is discovered, it is important to let the survey firm know that this is not acceptable and the data collection must be done again to ensure high standards. To reduce and detect these cases, make sure an independent auditing team is in place to oversee the data collection. It is customary to audit 10–15 percent of surveys to ensure that respondents exist and that data was collected accurately. When problems are found, some enumerators may need to be retrained or even fired.

Data loss. This can happen if completed questionnaires are lost or computers are stolen or malfunction. Computer data should always be backed up. In the field, surveys should be collected as soon as possible from interviewers, two to three times per week, if possible, to protect against loss. Should data be completely lost, it is best to go back and recollect data. This means revisiting individuals already surveyed and explaining to them that we need to ask the questions again. This can be very annoying to the respondents and costly for the program.

Data entry. Data entry should be performed promptly as surveys are collected. This allows problems to be identified and corrected in the field quickly. In addition, errors often occur during data entry. Most data entry computer packages allow for (but do not require) double entry, in which each value must be entered twice. Transcription errors are further minimized by the use of mobile phones, PDAs, laptop computers, or tablets in data entry.

Wrong assumptions. The main assumptions for the chosen evaluation design may not hold. By always using verification and falsification tests (see appendix 3), we can detect these cases during baseline analysis and take accurate action, including modifying the evaluation strategy. To reduce the chances that our chosen design is invalidated, it is important that the evaluation and program staff maintain close communication and cooperation, ensuring that program registration and data collection are in line with the evaluation requirements.

Conducting a Follow-up Survey and Analysis

Attrition. Attrition is a big problem for studies and can greatly decrease the value of the findings. Clearly, prevention is better than mitigation. Obtaining good contact information during baseline, providing incentives for youth to participate in the survey, and using tracking surveys can help minimize attrition. If, despite prevention efforts, the program experiences high attrition, one mitigation technique is to select a random sample of individual who have not been located and to conduct a very aggressive search for them. These individuals, if found, may adequately represent those not tracked. Finally, since some attrition is unavoidable, it is also possible to account for that attrition when defining the evaluation sample. Making the sample 10–20 percent bigger than it would need to be allows for a large enough number of survey responses to find statistically significant results even given attrition (though this does not offset the potential bias from attrition).

Noncompliance. In addition to attrition, there may be other cases where people

do not fully comply with a program's selection criteria. For example, youth selected to participate in a training program may actually not attend, while others who were assigned to the comparison group may actually be attend. A strict comparison of outcomes between the official treatment group and the comparison group will then misrepresent the actual impact of the program. As long as these cases are limited, and we know who exactly in the treatment and comparison groups received how much training (via program records), it is possible to correct for noncompliance using statistical techniques, the "treatment-on-the-treated" estimate, which the evaluator will be able to calculate.

Black-box evaluation. Another common problem at follow up is the lack of knowledge about how well the program was implemented. This leads to evaluations that cannot attribute observed changes (or the lack thereof) to program design or implementation. A common solution is to integrate findings from the monitoring system and to complement the impact evaluation with a process evaluation (also see Mixed Methods in note 8).

Disseminating Findings

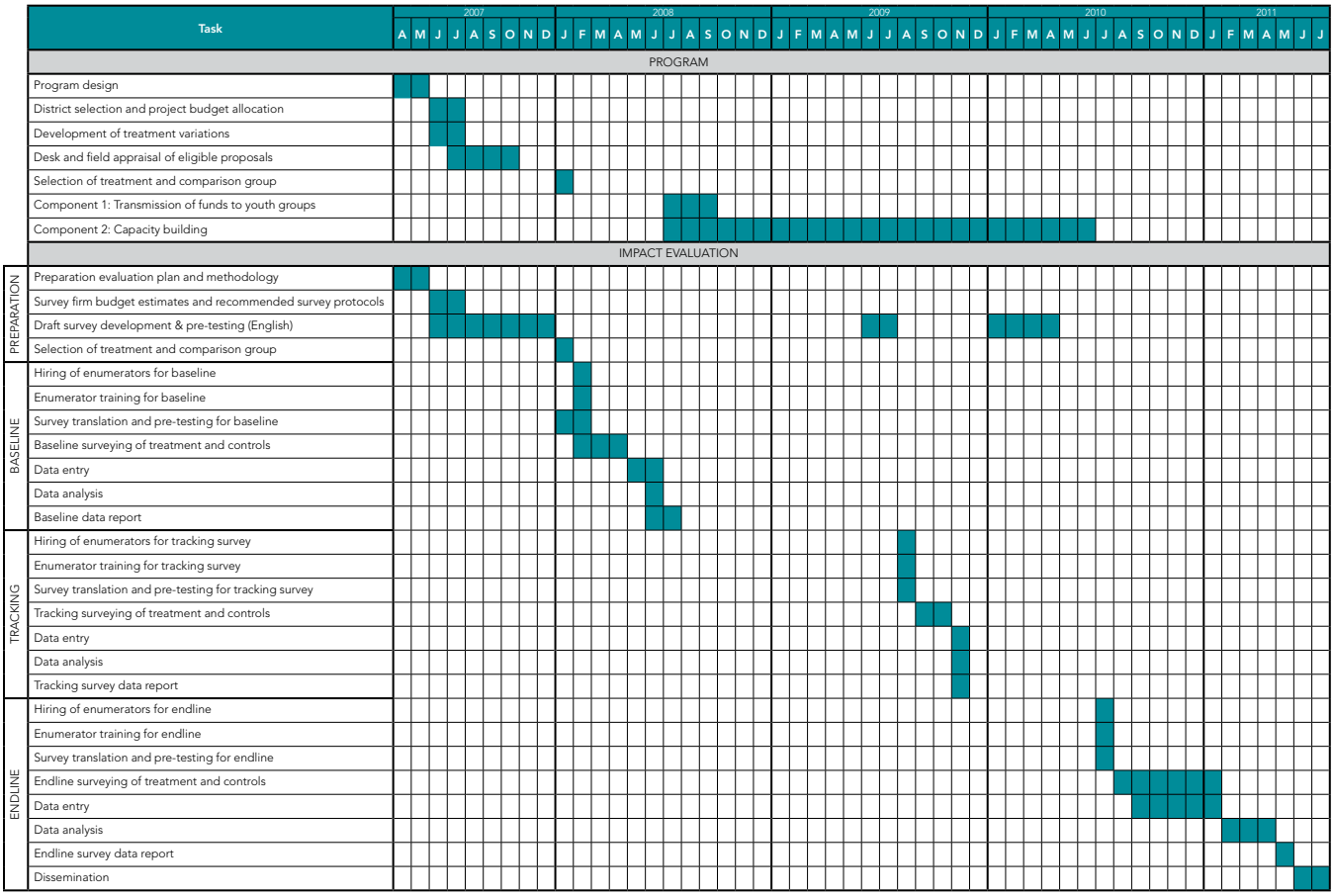
Limited use of the evaluation findings. If the results of the evaluation are not sufficiently shared with internal and external stakeholders, then the evaluation's main objectives of learning for the program and the youth livelihood sector at large are compromised. One way to overcome this issue is to define a dissemination strategy from the outset of the evaluation and to insist that at least one program staff work closely with the evaluation team. Thus, at least one key person in the program understands the evaluation and is well positioned to implement some of the findings of the report.

Key Points

1. Conducting an impact evaluation can be an expensive and time-consuming task, with many potential pitfalls. It is therefore essential to convene a high-quality team that can work on the evaluation over an extended period of time.
2. The evaluation plan is the first major product of an impact evaluation. It lays out the strategy for how to evaluate the intervention, including the research methodology, the sample size, the data collection plan, and other elements.
3. Interviewing children and youth poses particular challenges from obtaining parental consent to using appropriate language, so hiring a survey expert is advisable. Moreover, evaluations can raise ethical questions, so IRB approval should be sought for the evaluation design and the survey.
4. Conducting a baseline survey is highly recommended as it provides valuable information to inform program design and allows us to verify the feasibility of the chosen evaluation design.
5. The timing of the follow-up data collection has to be well thought through to capture the outcomes of interest, some of which may occur more in the short term, while others may need years to materialize.
6. It is crucial that evaluation findings, whether positive or negative, are widely disseminated. Sharing findings with internal, local, and international stakeholders provides the basis for learning and feedback.

NUSAF Case Study: Implementation of the Impact Evaluation

The NUSAF Youth Opportunities Program evaluation began in June 2007 and was completed in May 2011 with the development of the endline report. The program distributed funds to participants in August to September of 2008. The evaluation included a baseline survey in early 2008, a tracking survey in late 2009 and an endline survey in late 2010–early 2011. Each of the surveys covered the entire population of participants.



Source: Blattman, Fiala, and Martinez (2011).

Key Reading

Baker, J. 2000. *Evaluating the Impact of Development Projects on Poverty: A Handbook for Practitioners*. Washington, DC: The World Bank. (Chapter 2 is relevant to this note.) http://siteresources.worldbank.org/INTISPMA/Resources/handbook.pdf

Bamberger, M., Rugh, J., and Mabry, L. 2006. *Real World Evaluation: Working under Budget, Time, Data and Political Constraints.* Thousand Oaks: Sage Publications. (See chapters 3–8.) http://realworldevaluation.org/

Gertler, P., Martinez, S., Premand, P., Rawlings, L., and Vermeersch, C. 2011. *Impact Evaluation in Practice*. Washington, DC: The World Bank. (See chapters 10–13.) http://www.worldbank.org/ieinpractice

NOTE 8: Increasing the Relevance of the Impact Evaluation

Although a standard quantitative impact evaluation in and of itself can be of great value to our program and organization, there are a variety of options to further enhance the quality of the analysis and increase the relevance of the results. First, impact evaluations can answer a variety of research questions, including, but well beyond, average program effects. Second, when combined with qualitative tools, impact evaluations can be much more informative than when relying purely on quantitative methods. Finally, the results of an impact evaluation can be leveraged for further analysis, for example to weigh total program benefits with total program costs (through cost-benefit analysis). This final note provides a brief overview of these tools so practitioners can make the most of their impact evaluation.

Measuring a Variety of Impacts

The basic impact evaluation answers the question "Did the program work"; that is, did it affect the outcomes of interest as defined in our program and learning objectives? The question of whether the program as a whole had an impact is an important one, but it is by no means the only question we may want ask.

First, it may be useful to have a more nuanced picture of the program's actual impact. This includes obtaining a better understanding of the following questions:

- Do outcomes vary across different groups of beneficiaries (e.g., boys benefit, but girls do not)?
- What is the short-term versus the long-term impact of the intervention?
- Does the program have positive or negative spillover effects? Are there any intended or unintended outcomes beyond the actual target group?

Second, we may also be interested in testing crosscutting designs (CCDs), testing how the effectiveness of our program changes as we modify the design. CCDs investigate the following questions:

- Is one program design more effective than another? We may want to compare alternative interventions (providing start-up grants versus start-up loans for young entrepreneurs, for example), or test the most effective combination of program components (training alone, training plus internship, and training plus internship and mentoring).
- What is the most effective dosage of program activities? For example, should we provide 20, 50, or 100 hours of training?

If properly designed, impact evaluations can provide answers to these questions, though it will be difficult to answer all questions with a single impact evaluation. Because each intervention will have different priorities and learning objectives, we can design the impact evaluation to answer the questions most relevant to our program. By addressing a broader set of questions, we can improve the relevance of the evaluation findings. Yet, it is also important to understand that additional data are required to evaluate elaborate questions and crosscutting designs (see table 8.1).

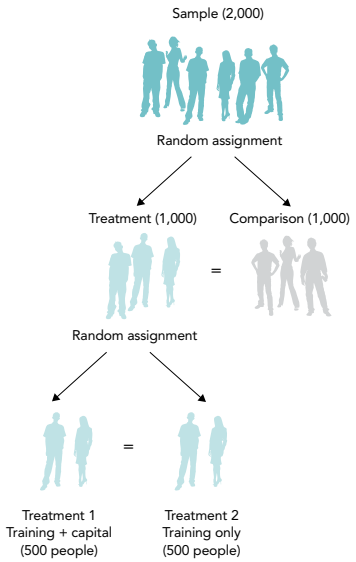
TABLE 8.1 Categories of impact evaluation questions

Question	Description	Additional Data Requirements	Sample Evaluation Result and Interpretation
What is the overall program impact on outcomes A, B, and C in group X? In context Y?	This is the standard impact evaluation question.	• n/a (standard data collection based on the method chosen)	The average impact of the training program on the income of youth is +$20 per month. The program has a positive impact on income.
Do the outcomes vary across population groups?	Interventions often affect groups differently (heterogeneity of impacts). Measuring only average impact may hide these differences, so we need to break down impacts by population group.	• Sociodemographic information of participants and comparison group (age, gender, income level, etc.) • To be able to disaggregate the results, the number of people covered by the evaluation (the sample size) needs to increase with each category of information that is to be analyzed	The average increase in income is $40 for boys, and $0 for girls. Older youth benefit more than younger youth ($30 versus $10 on average). Therefore, the program is not equally effective for all participants. We need to understand why groups benefit to a different extent and possibly adapt the program's targeting and design to accommodate particular groups.
What is the short-term versus the long-term impact of the program?	The change in outcomes may not be constant over time. Short-term effects may vanish, while long-term effects may not be visible for years after the intervention ended.	• Data over an extended period of time (in practice, it often means following treatment and comparison groups for several years)	At the end of the program, we observe an average monthly income for participants of -$5 (a loss) compared with the controls. Two years after the program, the average increase in monthly income for the treatment group is $20. Those who participated in the training were not able to work as much as their peers during the training, so they lost income. Over time, however, the training paid off and participants were able to secure incomes higher than those of their counterparts who did not participate. Looking only at short-term outcomes may provide misleading results.
Does the program have spillover effects?	The program may have indirect effects on nonparticipants (positive and negative).	• Data beyond the treatment and comparison group, to include family or community members	Not only do participants have a $20 higher average income, their neighbors also experienced a $5 increase. Participants apparently passed on the knowledge to others.
Is program design A or program design B more effective?	There is often ambiguity about the best possible program design. Questions can relate to comparing alternative interventions or combinations of program components.	• Several treatment groups (one receives design A, one receives design B, etc.) • The number of people covered by the evaluation needs to be large enough to be able to create more than one treatment group as well as a comparison group.	The average increase in income is $5 for those who received training and $30 for those who received training and an internship. Thus, providing practical work experience in addition to training appears to significantly improve impact.
What is the most effective dosage of the intervention?	More is not always better; finding the right balance of how much service to provide is important to maximize impact on the one hand and minimize costs on the other.	• Several treatment groups (one receives design A, one receives design B, etc.) • The number of people covered by the evaluation needs to be large enough to be able to create more than one treatment group as well as a comparison group.	The average increase in income is $0 for those who received one month of training, $20 for those who received three months, and $20 for those who received six months. Although 1 month of training was insufficient, six months of training had no additional benefit compared with three months of training. The optimal length of the training seems to be about three months.

CCDs help identify more than just the overall program impact; they also evaluate specific program features and why these do or do not work. For example, a program may provide vocational and entrepreneurial skills training, such as carpentry or tailoring, along with small start-up capital for businesses. The provision of cash grants could be expensive or politically difficult, and so the program director may wonder if the start-up capital is necessary, or if participants are able to implement their training without the capital. A CCD can help determine the best program design in this case.

CCDs require at least two treatment groups that receive different combinations or dosages of the program. These two groups can then be compared at the endline, and the difference between the two groups is the impact of the specific design. Using the example above, the program may conduct an evaluation in which a sample of 2,000 participants is randomly assigned into treatment and comparison groups. The treatment group can then be further randomized into two treatments. In treatment 1, the training and start-up capital are provided. In treatment 2, only the training is given (see figure 8.1).

FIGURE 8.1 Outline of an impact evaluation with a crosscutting design component



The impact of providing start-up capital can then be determined by comparing those in treatment 1 to those in treatment 2 at the endline. The impact of the training is determined by comparing those in treatment 2 to those in the comparison group.

Using Mixed-Methods Approaches

It is important to keep in mind the limitations of quantitative impact evaluation methods. If used in isolation, there is a risk that we will not be able to understand the complexity of program results and adequately interpret the impacts that may be identified. In order to have a solid understanding of the dynamics of an intervention and to be able to explain why things may be working, it is important that impact evaluation techniques are embedded in a framework of strong monitoring and process evaluation. Overall, we believe that using mixed methods—that is, explicitly adopting both quantitative and qualitative methods in the impact evaluation design—can significantly improve the learning in and about our programs.

As Bamberger, Rao, and Woolcock (2010, pp. 6–7) and Leeuw and Vaessen (2009) point out, there are several ways in which mixed methods can strengthen quantitative impact evaluation:

- Quantitative impact evaluations usually do not collect information on the quality of program implementation. Understanding the implementation process is crucial to understanding how program implementation affected program results and to correctly interpreting findings to differentiate whether disappointing results are due to weaknesses in program design or in implementation. Solid monitoring is therefore a prerequisite for effective evaluation and can be complemented with additional process analysis tools such as key informant interviews, direct participant observation, and focus groups.
- Incorporating qualitative methods can aid understanding of how and why the effect of the intervention may have varied across the target populations. Even though quantitative techniques can be designed to capture impact heterogeneity across groups, they cannot provide a clear understanding why these heterogeneities may have occurred.
- Although quantitative designs alone may be unable to capture the range of local circumstances in which each program is implemented, mixed methods can help provide detailed contextual analysis and document differences in the quality or speed of implementation across program sites. This qualitative information, in turn, can explain the potential differences in the outcomes of programs in different geographic areas.
- Many outcomes of youth livelihood interventions (such as mental health, empowerment, or household relations) are complex and multidimensional and may not be captured with quantitative methods. Mixed methods allow for tracking qualitative indicators and provide selected case-study analysis to help better understand the dynamics and results of the intervention. For example, small structured and semi-structured qualitative interviews in which participants are free to express real-life stories that fall outside categories of quantifiable information can help round out an understanding of a program's impact. Qualitative methods may also be better suited for collecting information on sensitive topics, such as reproductive health or violence.
- Qualitative methods may help identify appropriate indicators in the first place. For example, a focus group may yield important information about beneficiary concerns and how they expect the intervention to affect their lives.

Practically speaking, incorporating qualitative elements into our impact evaluation can take many forms, including open-ended survey questions, selected in-depth interviews and case studies, focus group discussions, participatory tools like the "Most Significant Change" technique (see Davies and Dart 2005), participant observation, and the like. To learn more about participatory monitoring and evaluation, consult Catley and colleagues (2010), Sabo Flores (2008), Powers and Tiffany (2006), and Gawler (2005).

At the same time, qualitative data alone are not well suited to identify program impacts. Using mixed methods, therefore, allows us to combine the strengths and offset the weaknesses of both qualitative and quantitative evaluation tools, allowing for an

overall stronger evaluation design (box 8.1 provides an example). In fact, the combined use of several research methods increases the credibility and validity of our results.

It is important to note, however, that using mixed-method designs can involve additional costs, time, and logistical challenges. In addition, it is often the case that the professional divisions among disciplines and researchers can make "building a multidisciplinary team time consuming and challenging" (Bamberger, Rao, Woolcock 2010, p. 17).

BOX 8.1 Example of mixed method evaluation

In an evaluation of Junior Achievement's (JA) *Our Nation* curriculum, evaluators combined a range of quantitative and qualitative research methods. *Our Nation* is one of several JA Worldwide globally distributed programs for elementary schools and consists of a series of lessons for students aged 9–11 that examine issues related to entrepreneurship, resources needed for business, and globalization. On the one hand, the evaluation relied on an experimental design with random assignment of students to treatment and comparison groups. Comparison students were from the same states and regions as the treatment students but their classes were randomly assigned to receive the program after the evaluation was completed. Moreover, the evaluators conducted several case studies, using teacher, volunteer, and JA staff interviews, student focus groups, and classroom observation.

The quantitative evaluation results demonstrated some positive impacts on students' content knowledge related to entrepreneurship and globalization but no effects on levels of school engagement and the acquisition of 21st century skills. In addition, the qualitative tools allowed for an in-depth understanding of the mechanisms at work. On the one hand, qualitative tools confirmed a good quality of implementation, with the majority of sessions being implemented according to JA guidance and high levels of satisfaction reported by students, teachers, and volunteers. However, they also indicated challenges to good program delivery, including, for example, insufficient time for volunteers to cover all the contents. Finally, the qualitative evaluation results suggested ways to improve the program, including extending time for sessions, reducing the difficulty of the vocabulary, and providing more teaching guidance for volunteers.

Source: RMC Research (2009).

Cost-Benefit and Cost-Effectiveness Analyses

In many cases, organizations may use different strategies to tackle the same problem. For example, to increase employability, we may want to improve career counseling or improve training. Even when a single strategy is pursued, we may take different approaches to implementation, such as using either public or private training providers. If these various programmatic or implementation strategies were shown to have the same impact, for example, if they each were shown to improve the probability of employment three months after the intervention by 50 percent, would we be equally happy implementing one approach over the other? Probably not. It is not enough to know that an intervention works, for whom, and in what context; we also need to know at what cost.

Having a realistic estimate of the costs, in turn, allows us to answer the following questions:

- How can we choose among alternatives? Which program is the most cost-efficient given a certain level of impact?

- Would we be able to scale up? If costs are high, it is unlikely that we will be able to reach a large number of beneficiaries.
- Is any intervention always better than none? If the total costs outweigh the total benefits of the program, maybe the resources are better spent somewhere else.

Analytical Tools

The two tools commonly used to answer the above questions are cost-effectiveness analysis (CEA) and cost-benefit analysis (CBA):

Cost-effectiveness analysis identifies the full cost of a program and relates these costs to specific measures of outputs or outcomes ($500 per person trained, per job created, per HIV/AIDS infection prevented, and the like). CEA thus tells us how much output or outcome we get per dollar spent, thereby identifying the most efficient allocation of resources when we compare alternative programs against the same criterion (see table 8.2).

TABLE 8.2 Cost-effectiveness estimates for *Jóvenes* programs

Country	Program	Cost Per Participant (in 2005 US$)
Argentina	Proyecto Joven	$1159
Chile	Chile Joven	$825–$1051
Peru	PROJoven	$697

Source: Betcherman et al. (2007).

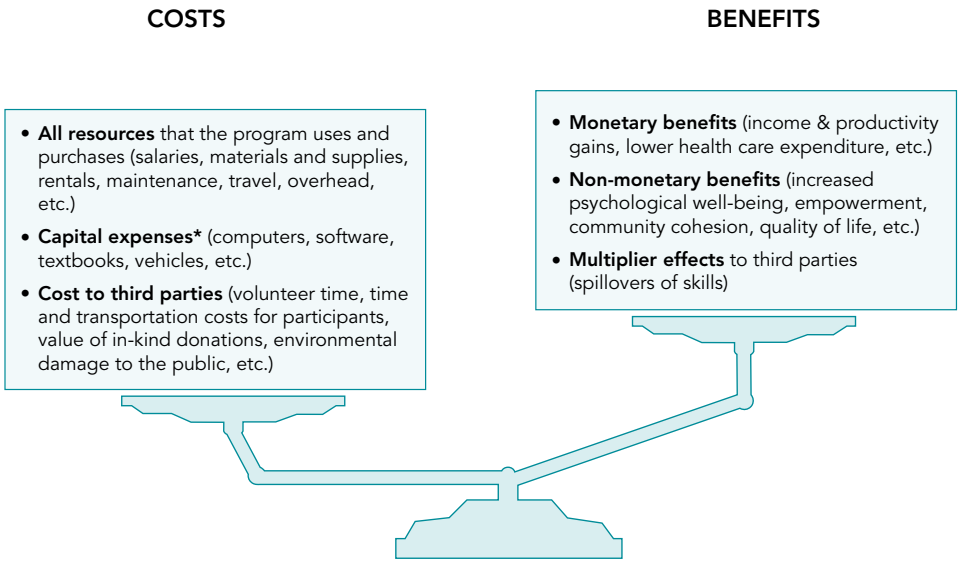
Cost-benefit analysis also identifies and quantifies the full cost of a program and further weighs those costs against the dollar value of all program benefits. Knowing the net benefits and net costs of the intervention, it is then possible to calculate the ratio of benefits to costs and to determine the return to society on the organization's investment. For example, the benefits:cost ratio is 2:1 if net benefits are $1,000 per person and net costs are $500. Overall, CBA seeks to determine whether benefits outweigh costs; that is, whether society is richer or poorer after making that investment.

Both CEA and CBA can be used before the intervention or during or after the program. However, only retrospective analysis will provide practitioners with the full information of actual costs and benefits to determine the overall success of the intervention. In fact, an impact evaluation is a necessary condition for having a reliable estimate of the program's direct and indirect benefits.

Capturing All Benefits and Costs

Cost-effectiveness and cost-benefit analyses require capturing, quantifying, and comparing all known costs (and, for CBA, known benefits) of the program to everyone directly or indirectly affected by the intervention: the implementing organization, the program beneficiaries, the government, and others (see figure 8.2).

FIGURE 8.2 Weighing costs and benefits



*Purchase of materials whose use exceeds one year.

Social return on investment (SROI) is variation of CBA that compares extra-financial benefits relative to the resources invested. It assigns financial proxy values to all those outcomes identified by stakeholders that do not typically have market values. To learn more, please consult the SROI Network's Web site: http://www.thesroinetwork.org/

Trying to put a dollar value on many intangible benefits may be difficult and subjective, and it can represent a big challenge, especially for CBA. Hence, CBA is usually considered most useful when there are multiple types of benefits and consensus about how to quantify them in monetary terms (J-PAL 2011).

Calculating Net Benefits for Participants

The benefits of a program can be measured by its impact on individual participants. For example, say a skills training program is found to increase income by $100 per person, per year, on average. In many areas in sub-Saharan Africa, this is a significant amount of money, and may represent great success for the program.

Assume the program costs $1 million to implement, with $400,000 to conduct the training and $600,000 in overhead, including all staff salaries. If it reaches 1,000 people, the program thus costs $1,000 per person to implement, with $400 going toward training and $600 going toward overhead. Is it worth running?

The answer to this question is based on three criteria.

First, **the program's impact must equal or exceed the impact of giving individuals cash equal to the cost of running the program**. In the example above, the impact must be compared with the effects of giving each person $1,000 cash. There are two possible scenarios. First, a person actually uses the $1000 and purchases training with that money. The cost of the training is still only $400 per person, which yields the same $100 per person per year return. The individual then has an extra $600 to use how they please, and is thus better off than with the program. In a second scenario, a person may use the money for something other than training that is less useful for her over the long term, such as cigarettes or other nonessential consumer goods. In the latter case, the program is worth running.

Second, **the program must have equal or greater return than running other programs**. Is it possible that another program could have realized the same or greater

impact per person for less money? This question requires a comparison of results across different program options. The program that has the greatest impact but costs the least is then the best program to continue running.

Finally, **the net present value of the return should be more than the cost of the program**. Present value is a way of thinking about the value of money today compared with its value in the future. Using the current example, we take the value of the money obtained yearly (the $100 per person return on training) and adjust its value over a period of time according to the *discount rate*, which in most cases equals, the local interest rate. The net present value is simply the sum of the present value adjusted over a period of time. This is represented in table 8.3, using an interest rate of 20 percent, which is a common rate in many developing countries.

TABLE 8.3 Present value and net present value for a yearly return of $100

Year	Present Value
0	$100
1	$83
2	$69
3	$58
4	$48
5	$40
6	$33
7	$28
8	$23
9	$19
10	$16
Net present value	$517

In this example, today $100 is valued at $100. However, at our current interest rate, the $100 of income today will be worth only $16 in ten years. Over a 10-year period, the net present value of our $100 is $517. Over the entire lifespan of a participant, the net present value will be at most $600. Thus, a return of $100 per person per year works out to a maximum return of $600 per person over their lifetime.

According to these criteria, in today's dollars, our outlay for the program ($1,000) is greater than the benefit to individuals, even though the cost for training ($400) is less than the net present value of the training ($517). Thus, unless people can be induced to take up the training on their own without the need for the overhead budget, or unless the overhead budget can be greatly reduced, the value of the training is not enough to justify the program.

Calculating Net Benefits for Society

In order to assess the net benefits to society we need to consider spillover effects. Spillovers refer to the positive or negative impacts the program has on those who are not directly involved with the program. There are two types of spillover effects that concern us here: multiplier effects, and prize and quantity effects.

Multiplier Effects

Multiplier effects occur when participants in a program impart their skills to others who were not formally associated with the program. For instance, a man trained in carpentry may train his son-in-law. The impact evaluation may measure only the impact the program had on the participant; it may miss the effect the program had on the son-in-law.

Using the example above, the impact on the carpenter is $100 per year. The impact on the son-in-law may be smaller due to lower quality training, but clearly the training of one person has improved the livelihood prospects for two people. The cost per person is thus lower than the $1,000 originally calculated.

Indirect benefits may be significant and could justify the costs of a program in some cases. In order to capture these spillover effects, plans should be made during the endline data collection to ask about others who have received training or otherwise benefited from the program.

Prize and Quantity Effects

Even though interventions may target only certain aspects of the population and local market, they can have effects on the larger economy, often referred to as "general equilibrium" effects. For example, if, as a result of our carpentry training, there are additional skilled carpenters in the local economy, competition among them may decrease prices for consumers. It is also possible that a program has negative spillovers on the economy. For instance, introducing extra tailors in an area where there are already a lot of tailors may drive prices so low that some of the tailors go out of business. This effect could significantly dilute the impact of a program.

Another undesired effect may result from negative consequences for nonparticipants. If participants of a particular intervention obtain a competitive edge in the labor market, for example, this may result in other youth not finding a job even though they would have in the absence of the program. Such effects are commonly referred to as displacement effects.

For most programs, prize and quantity effects are likely to be very small and not worth collecting data on. Large programs may wish to explore ways to capture this with an evaluation expert. One possibility may be to randomize the intervention at the community or district level.

Key Points

1. Depending on the learning objectives of a program and organization, it is worth exploring whether an impact evaluation can be designed to measure more than just the average impact of the program. Such additional impact questions can relate to heterogeneity, time-horizon, spillover effects, or the relative effectiveness of different program design options.
2. It is highly advisable to incorporate qualitative research elements into an impact evaluation. Using mixed-methods gives us a more comprehensive and nuanced

understanding of a program's impact, or lack thereof.

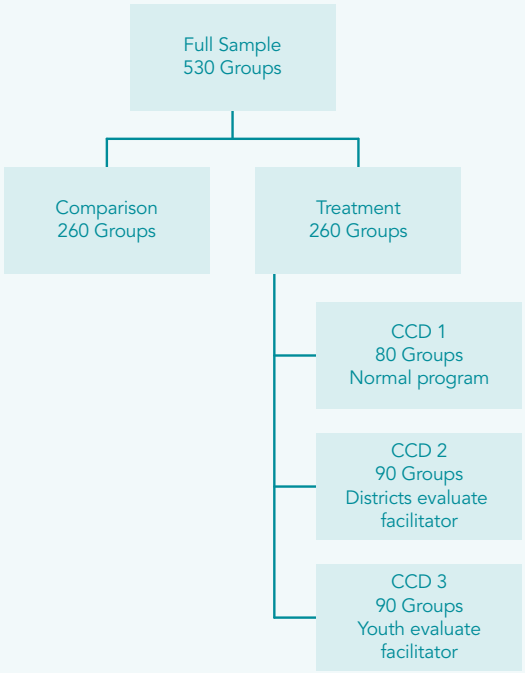
3. Information about the impact of a program may be of limited usefulness unless we also know the costs of designing and implementing the intervention. Any scale up will depend on this piece of information. It is therefore desirable to complement an impact evaluation with a cost-effectiveness or cost-benefit analysis.

NUSAF Case Study: Increasing the Relevance of the IE

Crosscutting Design

In addition to evaluating the overall effects of the Youth Opportunities Program, the impact evaluation was leveraged to test a complementary pilot intervention on an innovative program design variant. Anecdotal evidence from previous rounds of program funding suggested that the quality management, planning, and extension services provided by the district and the community facilitators are key determinants of individual youth group success. The impact evaluation therefore wanted to assess the effectiveness of giving an additional payment to hire a monitoring and extension advisor (MEA) that would be chosen by the group of youths themselves.

The treatment groups were randomly assigned to participate in the crosscutting design. Funded projects were randomly assigned to one of three groups. Treatment group 1 is treated as normal, with no additional intervention. Treatment group 2 has the district officers evaluate the MEAs, and treatment group 3 was given additional resources and asked to evaluate the MEAs themselves.



(continued)

NUSAF Case Study: Increasing the Relevance of the IE (cont'd)

Mixed Methods

The Youth Opportunities Program evaluation took advantage of quantitative and qualitative questions. The quantitative questionnaire was administered to approximately 2,600 youth, while the qualitative questionnaire was administered to about 100 youth. The qualitative questions included the following categories of interest:

1. Quality of group dynamics and cooperation, including process of group formation; group leadership and structures; group decision-making processes; past, present, and future of group activities; benefit and challenges of working in groups; and individual reasons for choosing to work in groups despite challenges.
2. NUSAF funds allocation, including group processes of fund allocation, group funding priorities versus project original plans, and deviation and other unofficial uses of fund.
3. Training experience, including process of choosing group skills training, confidence to apply skills learned, and benefits and challenges of applying skills as a livelihood strategy.
4. Livelihood strategies, including building livelihood after vocational training; risks, success, and failure associated with new livelihood strategies; reasons for success or failure; alternative livelihood strategies, and other strategies to deal with risks and shocks.
5. Empowerment and community participation, including sense of belonging in the communities, civic participation, gender relations, social support, social barriers, and relations with neighbors.

Quantitative data cannot alone bring out the full richness of a program. The responses to these questions will be used to better understand how NUSAF changed the lives of participants, as well as to provide some stories to help understand how the program impacted lives.

Source: Blattman, Fiala, and Martinez (2011).

Key Reading

Bamberger, M., Rao, V., and Woolcock, M. 2010. *Using Mixed Methods in Monitoring and Evaluation: Experiences from International Development*. Policy Research Working Paper 5245, Washington, DC: The World Bank. http://www-wds.worldbank.org/servlet/WDSContentServer/WDSP/IB/2010/03/23/000158349_20100323100628/Rendered/PDF/WPS5245.pdf

Bamberger, M., Rugh, J., and Mabry, L. 2006. *Real World Evaluation: Working under Budget, Time, Data and Political Constraints*. Thousand Oaks: Sage Publications. (See chapter 13.) http://realworldevaluation.org/

Cellini, S. R., and Kee, J. E. 2010. "Cost-Effectiveness and Cost-Benefit Analysis." In: Wholey, J., Hatry, H. P., and Newcomer, K. E., eds. *Handbook of Practical Program Evaluation*, 3rd ed. San Francisco: Jossey-Bass. http://home.gwu.edu/~scellini/CelliniKee21.pdf

Knowles, J., and Behrman, J. 2005. *A Practical Guide to Economic Analysis of Youth Projects*. HNP Discussion Paper. Washington, DC: The World Bank. http://siteresources.worldbank.org/HEALTHNUTRITIONANDPOPULATION/Resources/281627-1095698140167/KnowlesPracticalGuideFinal.pdf

Notes

Works Cited

Almeida, R., and Galasso, E. 2008. *Jumpstarting self-employment? Evidence from welfare participants in Argentina*. Washington, DC: The World Bank. http://ftp.iza.org/dp2902.pdf

Asian Development Bank. 2007. *Handbook on Social Analysis, A Working Document*. Manila: ADB. http://www.adb.org/Documents/Handbooks/social-analysis/Appendixes.pdf

Attanasio, O., Kugler, A., and Meghir, C. 2009. "Subsidizing Vocational Training for Disadvantaged Youth in Developing Countries: Evidence from a Randomized Trial." IZA Discussion Paper No. 4251. Bonn: IZA. http://ftp.iza.org/dp4251.pdf

Baker, J. 2000. *Evaluating the Impact of Development Projects on Poverty: A Handbook for Practitioners*. Washington, DC: The World Bank. http://siteresources.worldbank.org/INTISPMA/Resources/handbook.pdf

Bamberger, M., Rao, V., and Woolcock, M. 2010. "Using Mixed Methods in Monitoring and Evaluation: Experiences from International Development." Policy Research Working Paper No. 5245. Washington, DC: The World Bank. http://www-wds.worldbank.org/servlet/WDSContentServer/WDSP/IB/2010/03/23/000158349_20100323100628/Rendered/PDF/WPS5245.pdf

Beauvy-Sany, M., Conklin, S., Hershkowitz, A., Rajkotia, R., and, Berg, C. 2009. *Guidelines and Experiences for Including Youth in Market Assessments for Stronger Youth Workforce Development Programs*. Washington, DC: The SEEP Network. http://www.seepnetwork.org/Resources/YouthPLP_Assessments.pdf

Bertrand, A., Beauvy-Sany, M., Cilimkovic, S., Conklin, S., and Jahic, S. 2009. *Monitoring and Evaluation for Youth Workforce Development Projects*. Washington DC: The SEEP Network. http://seepnetwork.org/Resources/YouthPLP_MonitoringEval.pdf

Betcherman, G., Godfrey, M., Puerto, S., Rother, F., and Stavreska, A. 2007. "A Review of Interventions to Support Young Workers: Findings of the Youth Employment Inventory." SP Discussion Paper No. 0715. Washington, DC: The World Bank. http://www.youth-employment-inventory.org/downloads/1.pdf

Blattman, C., Fiala, N., and Martinez, S. 2011. *Employment Generation in Rural Africa: Mid-term Results from an Experimental Evaluation of Youth Opportunities Program in Northern Uganda*. New Haven, CT: Innovations for Poverty Action. https://www.poverty-action.org/sites/default/files/blattmanfialamartinez.midtermreport.pdf

Bose, R. 2010. "A Checklist for the Reporting of Randomized Control Trials of Social and Economic Policy Interventions in Developing countries: CEDE Version 1.0." Working Paper No. 6. New Delhi: International Initiative for Impact Evaluation. http://www.3ieimpact.org/admin/pdfs_papers/working paper 6.pdf

Brady, M., Salem, A., and Zibani, N. 2007. "Bringing New Opportunities to Adolescent Girls in Socially Conservative settings: The Ishraq Program in Rural Upper Egypt." *Transitions to Adulthood*, Brief No. 12. New York: Population Council. http://www.popcouncil.org/pdfs/IshraqFullReport.pdf

Brener, N. D., Billy, J. O .G., and Grady, W.R. 2003. "Assessment of Factors Affecting the Validity of Self-Reported Health-Risk Behavior Among Adolescents: Evidence From the Scientific Literature." *Journal of Adolescent Health* 33: 436–457. http://www.cdc.gov/HealthyYouth/yrbs/pdf/validity.pdf

Bronfenbrenner, U. 1979. *The Ecology of Human Development: Experiments by Nature and Design*. Cambridge, MA: Harvard University Press. http://books.google.com/books?id=OCmbzWka6xUC&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

Bruhn, M., and Zia, B. 2011. "Stimulating Managerial Capital in Emerging Markets: The Impact of Business and Financial Literacy for Young Entrepreneurs." Policy Research Working Paper No. 5642. Washington, DC: The World Bank. http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2011/04/27/000158349_20110427082512/Rendered/PDF/WPS5642.pdf

Card, D., Kluve, J., and Weber, R. 2009. "Active Labor Market Policy Evaluations: A Meta-Analysis." IZA Discussion Paper No. 4002. Bonn. http://ftp.iza.org/dp4002.pdf

Cooley, L. 1989. "The Logical Framework: Program Design for Program Results." *The Entrepreneurial Economy Review* 8(1): 8–15. http://evaluation.zunia.org/post/the-logical-framework-program-design-for-program-results/

Creswell, John. 2008. *Research Design: Qualitative, Quantitative and Mixed Methods Approaches*. Thousand Oaks, CA: Sage. http://books.google.com/books?id=bttwENORfhgC&printsec=frontcover&source=gbs_ge_summary_r&cad=0

Cunningham, W., Sanchez-Puerta, M. L., and Wuermli, A. 2010. "Active Labor Market Policies for Youth: A Framework to Guide Youth Employment Interventions." Employment Policy Primer No. 16. Washington, DC: The World Bank. http://siteresources.worldbank.org/INTLM/214578-1103128720951/22795057/EPPNoteNo16_Eng.pdf

Davies, R. J., and Dart, J. 2005. *The "Most Significant Change" (MSC) Technique: A Guide to Its Use*. http://mande.co.uk/docs/MSCGuide.pdf

Development Marketplace. 2008. *Development Marketplace Grantee Toolkit: Project Design, Monitoring and Evaluation for Small Innovative Projects*. Washington DC: The World Bank. http://siteresources.worldbank.org/DEVMARKETPLACE/Resources/DM_Grantee_Toolkit_Final.pdf

DFID (Department for International Development). 1999. "Framework." Sustainable Livelihood Guidance Sheets, Section 2. Oxford: DFID. http://www.eldis.org/vfile/upload/1/document/0901/section2.pdf

Duflo, E., Glennerster, R., and Kremer, M. 2006. "Using Randomization in Development Economics Research: A Toolkit." BREAD Working Paper No. 136. http://www.povertyactionlab.org/sites/default/files/documents/Using%20Randomization%20in%20Development%20Economics.pdf

Fiszbein, A., and Schady, N. 2009. *Conditional Cash Transfer, Reducing Present and Future Poverty*. Policy Research Report. Washington, DC: The World Bank. http://siteresources.worldbank.org/INTCCT/Resources/5757608-1234228266004/PRR-CCT_web_noembargo.pdf

GAO (U.S. Government Accountability Office). 1991. *Designing Evaluations*. Washington, DC: GAO. http://www.gao.gov/special.pubs/10_1_4.pdf

Gawler, M. 2005. *Useful Tools for Engaging Young People in Participatory Evaluation*. Geneva: UNICEF CEE/CIS. http://www.artemis-services.com/downloads/tools-for-participatory-evaluation.pdf

Gertler, P., Martinez, S., Premand, P., Rawlings, L., and Vermeersch, C. 2011. *Impact Evaluation in Practice*. Washington, DC: The World Bank. http://www.worldbank.org/ieinpractice

Gosparini, P., Russo, L., Sirtori, M., and Valmarana, C. 2004. *The Monitoring and Evaluation Manual of the NGOs of the Forum Solint*. Rome, Italy: CISP, COSV, COOPI, Intersos, Movimondo, and DRN. http://www.cosv.org/echotrain/materiale/0B_ITA/ECHOTrain_Documenti/ECHOTrain_Documenti_Manuali/ECHOTrain_Documenti_Manuali_SOLINT/Manuale M&E- Solint.pdf

Hempel, K. 2006. Erfolgskontrolle in der deutschen Entwicklungszusammenarbeit: Zentrale Herausforderungen unter besonderer Berücksichtigung der Transparenz gegenüber der Öffentlichkeit. Stuttgart: ibidem-Verlag.

Ibarrarán, P., and Rosas Shady, D. 2009. "Evaluating the Impact of Job Training Programmes in Latin America: Evidence from IDB-Funded Operations." *Journal of Development Effectiveness* 1(2): 195–216. http://www.iza.org/conference_files/ELMPDC2009/ibarraran_p4263.pdf

Imas, L., and Rist, R. 2009. *The Road to Results: Designing and Conducting Effective Development Evaluations*. Washington, DC: The World Bank. http://books.google.com/books?id=NEsg-BtinIsC&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

Innovations for Poverty Action. 2010. *IPA IRB Submission Guidelines and Application*. Internal Document. New Haven, CT: IPA.

Jaramillo, M., and Parodi, S. 2003. Jóvenes Emprendedores. Lima, Peru: Instituto Apoyo. http://www.grade.org.pe/download/pubs/MJ-SP-J%C3%B3venes%20emprendedores.pdf

Jones, N., Jones, H., Steer, L., and Datta, A. 2009. "Improving Impact Evaluation Production and Use." Working Paper No. 3000. London: Overseas Development Institute. http://www.odi.org.uk/resources/download/3177.pdf

J-PAL (Abdul Latif Jameel Poverty Action Lab). 2011. "Cost-benefit/Effectiveness/Comparison Analysis." J-PAL Web site. http://www.povertyactionlab.org/methodology/what-evaluation/cost-benefit/effectiveness/comparison-analyses. Accessed May 31, 2011.

Karlan, D. 2009. "Thoughts on Randomised Trials for Evaluation of Development: Presentation to the Cairo Evaluation Clinic." *The Journal of Development Effectiveness* 1(3): 237–242. http://www.3ieimpact.org/admin/pdfs_papers/50.pdf

Kellogg (W. K. Kellogg Foundation). 1998. *Evaluation Handbook*. Battle Creek, MI: W. K. Kellogg Foundation. http://www.wkkf.org/knowledge-center/resources/2010/W-K-Kellogg-Foundation-Evaluation-Handbook.aspx

———. 2004. *Logic Model Development Guide*. Battle Creek, MI: W. K. Kellogg Foundation. http://www.wkkf.org/knowledge-center/resources/2006/02/WK-Kellogg-Foundation-Logic-Model-Development-Guide.aspx

Khandker, S., Koolwal, G., and Samad, H. 2010. *Handbook on Impact Evaluation: Quantitative Methods and Practices*. Washington, DC: The World Bank. http://www-wds.worldbank.org/external/default/WDSContentServer/IW3P/IB/2009/12/10/000333037_20091210014322/Rendered/PDF/520990PUB0EPI1101Official0Use0Only1.pdf

Klinger, B., and Schuendeln, M. 2007. "Can Entrepreneurial Activity be Taught? Quasi-Experimental Evidence from Central America." Working Paper No. 153. Cambridge, MA: Harvard University Center for International Development. http://www.hks.harvard.edu/var/ezp_site/storage/fckeditor/file/pdfs/centers-programs/centers/cid/publications/faculty/wp/153.pdf

Kusek, J. Z., and Rist, R. C. 2004. *Ten Steps to a Results Based Monitoring and Evaluation System: A Handbook for Development Practitioners*. Washington, DC: The World Bank. http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2004/08/27/000160016_20040827154900/Rendered/PDF/296720PAPER0100steps.pdf

Leeuw, F., and Vaessen, J. 2009. *Impact Evaluations and Development: NONIE Guidance on Impact Evaluation*. Washington, DC: Network of Networks on Impact Evaluation. http://siteresources.worldbank.org/EXTOED/Resources/nonie_guidance.pdf

Muzi, S. 2011. "Summary Notes on Impact Evaluations in the Active Labor Market/Youth Employment Programs Cluster Supported by SIEF." Washington, DC: The World Bank.

NIH (National Institutes of Health). 2008. *Protecting Human Research Participants*. Bethesda: NIH Office of Extramural Research. http://phrp.nihtraining.com/users/PHRP.pdf

NORC (National Opinion Research Center). 2007. *Operational Plan for the Baseline Survey of "Mi Primer Empleo" Youth Employment Program, Honduras*. Chicago: NORC.

OECD (Organization for Economic Cooperation and Development). 1991. *Principles for Evaluation of Development Assistance*. Paris: Development Assistance Committee. http://www.oecd.org/dataoecd/31/12/2755284.pdf

———. (n.d.). DAC Criteria for Evaluating Development Assistance. http://www.oecd.org/document/22/0,2340,en_2649_34435_2086550_1_1_1_1,00.html

Osborn, D., and Gaebler, T. 1992. *Reinventing Government: How the Entrepreneurial Spirit is Transforming the Pubic Sector*. Reading, MA: Addison-Wesley.

Penrose-Buckley, C. 2007. "Annex 2: Rapid Assessment of Markets and Producers." In: *Producer Organisations: A Guide to Developing Collective Rural Enterprises*. Oxford: Oxfam GB http://www.oxfam.org.uk/resources/downloads/produorgs_book.pdf

Powers, J. L., and Tiffany, J. S. 2006. "Engaging Youth in Participatory Research and Evaluation." *Journal of Public Health Management and Practice*, November (Suppl.), S79-S87. http://www.health.state.ny.us/community/youth/development/docs/jphmp_s079-s087.pdf

Ravallion, M. 2008. *Evaluating Anti-Poverty Programs. Handbook of Development Economics*, Vol. 4, Schultz, T. P., and Strauss, J. (eds). Amsterdam: Elsevier. http://siteresources.worldbank.org/INTISPMA/Resources/383704-1130267506458/Evaluating_Antipoverty_Programs.pdf

Rawlings, L., and Rubio, G. 2005. "Evaluating the Impacts of Conditional Cash Transfer Programs." *World Bank Research Observer* 20(1): 29–55. http://www.crin.org/docs/Evaluating the Impact of Cash Transfer Programs.pdf

RMC (RMC Research Corporation). 2009. *"Our Nation" Evaluation*. Denver, CO: RMC. http://www.myja.org/programs/evaluation/reports/our_nation.pdf

Roberts, B., Caspi, A., and Moffitt, T. 2003. "Work Experiences and Personality Development in Young Adulthood." *Journal of Personality and Social Psychology* 84(3): 582–593. http://www.mendeley.com/research/work-experiences-and-personality-development-in-young-adulthood-1/

Robins, R., Fraley, R., Roberts, B., and Trzesniewski, K. 2001. "A Longitudinal Study of Personality Change in Young Adulthood." *Journal of Personality* 69(4): 617–640. http://internal.psychology.illinois.edu/~broberts/Robins,%20et%20al,%202001.pdf

Rubio, G. 2011. "The Design and Implementation of a Menu of Evaluations." PREM Notes No. 6, The Nuts and Bolts of M and E Systems. Washington, DC: The World Bank. http://siteresources.worldbank.org/INTPOVERTY/Resources/335642-1276521901256/premnoteME6.pdf

Rossiasco, P., O'Neil, F., Richardson, A., and Francis, P. 2010. "Youth and Employment In Post-Conflict Countries: The Psycho-Social Dimension." *Child and Youth Development Notes* (4)2. Washington DC: The World Bank. http://siteresources.worldbank.org/INTLM/Resources/390041-1319047943696/CYDN_No2_Psychosocial_Employment.pdf

Sabo Flores, K. 2008. *Youth Participatory Evaluation: Strategies for Engaging Young People*. San Francisco: Jossey-Bass. http://www.josseybass.com/WileyCDA/WileyTitle/productCd-0787983926.html

Savedoff , W. D., Levine, R., and Birdsall, N. 2006. *When Will We Ever Learn? Improving Lives through Impact Evaluation*. Washington, DC: Center for Global Development. http://www.cgdev.org/content/publications/detail/7973

Shapiro, J. 2003. "Monitoring and Evaluation." Washington, DC: CIVICUS. http://www.civicus.org/new/media/Monitoring and Evaluation.doc

Taylor-Powell, E. 2005. "Logic Models: A Framework for Program Planning and Evaluation." University of Wisconsin–Extension, Program Development and Evaluation, 2005. slides 15–16. http://www.uwex.edu/ces/pdande/evaluation/pdf/nutritionconf05.pdf. Accessed June 2, 2011.

UNICEF (United Nations Children's Fund). 1991. "A UNICEF Guide for Monitoring and Evaluation: Making a Difference?" New York: UNICEF. http://preval.org/documentos/00473.pdf

Trochim, W. M. K. 2006. "The Nonquivalent Groups Design." Web Center for Social Research Methods, Research Methods Knowledge Base. http://www.socialresearchmethods.net/kb/quasnegd.php. Accessed June 2, 2011.

World Bank. 2002. *A Sourcebook for Poverty Reduction Strategies*. Washington, DC: The World Bank. http://siteresources.worldbank.org/INTPRS1/Resources/383606-1205334112622/4943_annex_c.pdf

———. 2006. *World Development Report 2007: Development and the Next Generation*. Washington, DC: The World Bank. http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2006/09/13/000112742_20060913111024/Rendered/PDF/359990WDR0complete.pdf

———. 2007a. *Data for Impact Evaluation*. Doing Impact Evaluation No. 6. Washington, DC: The World Bank. http://siteresources.worldbank.org/INTISPMA/Resources/383704-1146752240884/Doing_ie_series_06.pdf

———. 2007b. *Impact Evaluation for Microfinance: Review of Methodological Issues*. Doing Impact Evaluation No. 7. Washington, DC: The World Bank. http://siteresources.worldbank.org/INTISPMA/Resources/383704-1146752240884/Doing_ie_series_07.pdf

———. 2007c. "Miles to Go: A Quest for an Operational Labor Market Paradigm for Developing Countries." Washington, DC: The World Bank. http://siteresources.worldbank.org/INTLM/Resources/390041-1212776476091/5078455-1267646113835/MILESThequestoperationalLMparadigm_Jan212008.pdf

———. 2008. Project Paper for a project in the amount of USD 2,793,706 to the Republic of Liberia for a proposed Economic Empowerment of Adolescent Girls Project, Report No. 44985-LR. Washington, DC: The World Bank.

———. 2009. *Institutionalizing Impact Evaluation within the Framework of a Monitoring and Evaluation System*. Washington, DC: The World Bank. http://siteresources.worldbank.org/EXTEVACAPDEV/Resources/4585672-1251461875432/inst_ie_framework_me.pdf

———. 2011. "Evaluation of the Use of Entertainment Education to Improve Financial Capability in South Africa." Concept Note (February). Washington, DC: The World Bank.

Resources

Youth Development/Employment Literature

Bidwell, K., Galbraith, C., et al. 2008. *Market Assessment Toolkit for Vocational Training Providers and Youth*. New York: Women's Commission for Refugee Women and Children and Columbia University School of International and Public Affairs. http://www.womensrefugeecommission.org/docs/ug_ysl_toolkit.pdf

Bronfenbrenner, U. 1979. *The Ecology of Human Development: Experiments by Nature and Design*. Cambridge, MA: Harvard University Press. http://books.google.com/books?id=OCmbzWka6xUC&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

Card, D., Ibarrarán, P., and Villa, J. M. 2011. "Building in an Evaluation Component for Active Labor Market Programs: A Practitioner's Guide." IZA Discussion Paper No. 6085. Bonn: IZA. http://ftp.iza.org/dp6085.pdf

Cunningham, W., Cohan, L., Naudeau, S., and McGinnis, L. 2008. *Supporting Youth at Risk: A Policy Toolkit for Middle-Income Countries*. Washington, DC: The World Bank. http://siteresources.worldbank.org/INTCY/Resources/395766-1187899515414/SupportingYouthAtRisk.pdf

Cunningham, W., Sanchez-Puerta, M. L., and Wuermli, A. 2010. "Active Labor Market Policies for Youth: A Framework to Guide Youth Employment Interventions." Employment Policy Primer No. 16. Washington, DC: The World Bank. http://siteresources.worldbank.org/INTLM/214578-1103128720951/22795057/EPPNoteNo16_Eng.pdf

International Labour Office. 2010. *Global Employment Trends for Youth: Special Issues on the Impact of the Global Economic Crisis on Youth*. Geneva: ILO. http://www.ilo.org/wcmsp5/groups/public/---ed_emp/---emp_elm/---trends/documents/publication/wcms_143349.pdf

James-Wilson, D. 2008. *Youth Livelihoods Development Program Guide*. Washington DC: EQUIP3. http://www.equip123.net/docs/e3-LivelihoodsGuide.pdf

Making Cents International. *State of the Field in Youth Enterprise Development and Livelihoods Development series*. Washington, DC: MCI. http://www.youtheconomicopportunities.org/media.asp

Van Adams, A. 2007. *The Role of Youth Skills in the Transition to Work: A Global Review*. Washington, DC: The World Bank. http://siteresources.worldbank.org/INTCY/Resources/395766-1187899515414/RoleofYouthSkills.pdf

Van der Geest, K. 2010. *Rural Youth Employment in Developing Countries: A Global View*. Rome: FAO. http://www.fao.org/docrep/012/al414e/al414e00.pdf

Women's Refugee Commission. 2009. *Building Livelihoods: A Field Manual for Practitioners in Humanitarian Settings*. New York: WRC. http://www.womensrefugeecommission.org/docs/livelihoods_manual.pdf

World Bank 2006. *World Development Report 2007: Development and the Next Generation*. Washington, DC: The World Bank. http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2006/09/13/000112742_20060913111024/Rendered/PDF/359990WDR0complete.pdf

Monitoring and Evaluation Literature

Bamberger, M., Rugh, J., and Mabry, L. 2006. *Real World Evaluation: Working under Budget, Time, Data and Political Constraints*. Thousand Oaks: Sage. http://www.realworldevaluation.org

Bertrand, A., Beauvy-Sany, M., Cilimkovic, S., Conklin, S., and Jahic, S. 2009. *Monitoring and Evaluation for Youth Workforce Development Projects*. Washington DC: The SEEP Network. http://seepnetwork.org/Resources/YouthPLP_MonitoringEval.pdf

Catley, A., Burns, J., Abebe, D., and Suji, O. 2010. *Participatory Impact Assessment, A Guide for Practitioners*. Medford, MA: Feinstein International Center. https://wikis.uit.tufts.edu/confluence/download/attachments/19924843/Part_Impact_10_21_08V2.pdf?version=1&modificationDate=1225200269000

Duflo, E., Glennerster, R., and Kremer, M. 2006. "Using Randomization in Development Economics Research: A Toolkit." BREAD Working Paper No. 136. http://www.povertyactionlab.org/sites/default/files/documents/Using%20Randomization%20in%20Development%20Economics.pdf

Gertler, P. Martinez, S., Premand, P., Rawlings, L., and Vermeersch, C. 2011. *Impact Evaluation in Practice*. Washington, DC: The World Bank. http://www.worldbank.org/ieinpractice

Imas, L., and Rist, R. 2009. *The Road to Results: Designing and Conducting Effective Development Evaluations*. Washington, DC: The World Bank. http://books.google.com/books?id=NEsg-BtinIsC&printsec=frontcover&source=gbs_ge_summary_r&cad=0

Karlan, D., and Appel, J. 2011. *More than Good Intentions: How a New Economics Is Helping to Solve Global Poverty*. Boston, MA: Dutton. http://books.google.com/books?id=JOMDsm5Gn9kC&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

Khandker, S., Koolwal, G., and Samad, H. 2010. *Handbook on Impact Evaluation: Quantitative Methods and Practices*. Washington, DC: The World Bank. http://www-wds.worldbank.org/external/default/WDSContentServer/IW3P/IB/2009/12/10/000333037_20091210014322/Rendered/PDF/520990PUB0EPI1101Official0Use0Only1.pdf

Kluve, J. 2011. *Measuring Employment Effects of Technical Cooperation Interventions: Some Methodological Guidelines*. Second revised edition. Eschborn: GIZ http://www2.gtz.de/wbf/4tDx9kw63gma/Methodenleitfaden.pdf

Knowles, J., and Behrman, J. 2005. *A Practical Guide to Economic Analysis of Youth Projects*. HNP Discussion Paper. Washington, DC: The World Bank. http://siteresources.worldbank.org/HEALTHNUTRITIONANDPOPULATION/Resources/281627-1095698140167/KnowlesPracticalGuideFinal.pdf

Kusek, J. Z., and Rist, R. C. 2004. *Ten Steps to a Results-Based Monitoring and Evaluation System: A Handbook for Development Practitioners*. Washington, DC: The World Bank. http://www.oecd.org/dataoecd/23/27/35281194.pdf

Leeuw, F., and Vaessen, J. 2009. *Impact Evaluations and Development: NONIE Guidance on Impact Evaluation*. Washington, DC: Network of Networks on Impact Evaluation. http://siteresources.worldbank.org/EXTOED/Resources/nonie_guidance.pdf

Powell, E., Jones, L., and Henert, E. 2003. *Enhancing Program Performance with Logic Models*. University of Wisconsin–Extension, Program Development and Evaluation. http://www1.uwex.edu/ces/lmcourse/

Ravallion, M. 2008. *Evaluating Anti-Poverty Programs. Handbook of Development Economics*, Vol. 4, Schultz, T. P., and Strauss, J. (eds). Amsterdam: Elsevier. http://siteresources.worldbank.org/INTISPMA/Resources/383704-1130267506458/Evaluating_Antipoverty_Programs.pdf

Wholey, J., Hatry, H. P., and Newcomer, K. E., eds. 2010. *Handbook of Practical Program Evaluation*, 3rd ed. San Francisco: Jossey-Bass. http://books.google.com/books?id=zntNhoO6gCUC&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

World Bank 2007. *Doing Impact Evaluation*. A series of fourteen associated reports. http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTPOVERTY/EXTISPMA/0,,contentMDK:22268536~pagePK:210058~piPK:210062~theSitePK:384329,00.html

Research Databases

EconPapers http://econpapers.repec.org/

Education Resources Information Center (ERIC) http://www.eric.ed.gov/

IDEAS http://ideas.repec.org/

IZA Discussion Papers http://www.iza.org/en/webcontent/publications/papers

J-STOR http://www.jstor.org/

National Bureau of Economic Research (NBER) http://www.nber.org/

PsychInfo http://www.apa.org/pubs/databases/psycinfo/index.aspx

Social Science Research Network
http://papers.ssrn.com/sol3/DisplayAbstractSearch.cfm

Academic Journals

Journal of Adolescence
http://www.elsevier.com/wps/find/journaldescription.cws_home/622849/description#description

Journal of Early Adolescence
http://jea.sagepub.com/

Journal of Research on Adolescence
http://www.wiley.com/bw/journal.asp?ref=1050-8392

Journal of Youth and Adolescence
http://www.springer.com/psychology/child+%26+school+psychology/journal/10964

Journal of Youth Development
www.nae4ha.org/directory/jyd/index.html

Journal of Youth Studies
http://taylorandfrancis.co.uk/journals/titles/13676261.asp

New Directions for Child and Adolescent Development
http://www3.interscience.wiley.com/journal/85511342/home

New Directions for Youth Development
http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1537-5781/homepage/ProductInformation.html

The Future of Children
http://futureofchildren.org/

Vulnerable Children and Youth Studies
http://www.informaworld.com/smpp/title~db=all~content=t724921266

Young: Nordic Journal of Youth Research
http://you.sagepub.com

Databases of Existing and Ongoing Impact Evaluations

3ie International Initiative for Impact Evaluation
http://www.3ieimpact.org/database_of_impact_evaluations.html

Abdul Latif Jameel Poverty Action Lab (JPAL)
http://www.povertyactionlab.org/search/apachesolr_search?filters=type:evaluation

Innovations for Poverty Action (IPA)
http://www.poverty-action.org/project-evaluations/search

World Bank DIME Initiative
http://web.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTDEVIMPEVAINI/0,,contentMDK:21553788~pagePK:64168445~piPK:64168309~theSitePK:3998212,00.html

Youth Employment Inventory
http://www.youth-employment-inventory.org/

Blogs

Aid Watch
http://aidwatchers.com/

Chris Blattman
http://chrisblattman.com/

Development that Works (IDB)
http://blogs.cross-cutting.org/desarrolloefectivo_en/

Find What Works
http://findwhatworks.wordpress.com/

Good Intentions Are Not Enough
http://goodintents.org/

Innovations for Poverty Action
http://www.poverty-action.org/blog

Larry Dershem, Save the Children
http://designmonitoringevaluation.blogspot.com/

Monitoring and Evaluation News
http://mande.co.uk/

World Bank Development Impact
http://blogs.worldbank.org/impactevaluations/

Capacity Building

ILO International Training Centre Course on M&E for Youth Employment Projects
http://www.itcilo.org

International Program for Development Evaluation Training (IPDET)
http://www.ipdet.org

J-PAL Course on Impact Evaluation
http://www.povertyactionlab.org/course

World Bank Impact Evaluation Workshops
http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTPOVERTY/EXTISPMA/0,,contentMDK:21754074~menuPK:384336~pagePK:148956~piPK:216618~theSitePK:384329,00.html

YEN Evaluation Clinics
http://yenclinic.groupsite.org

You can browse a large selection of self-learning, online, and onsite trainings on monitoring and evaluation (beyond impact evaluation) on the *My M&E* Web site
http://www.mymande.org/index.php?q=training_search&x=admin

Web Sites

Evalsed: The resource for the evaluation of socioeconomic development
http://ec.europa.eu/regional_policy/sources/docgener/evaluation/evalsed/

My M&E
http://www.mymande.org/

Web Center for Social Research Methods
http://www.socialresearchmethods.net/kb/design.php

APPENDIX 1. Sample Indicators for Youth Assessments

Status of Youth	
Category	**Selected Indicators**
Poverty	• Number and percentage of youth living on less than $2 per day • Number and percentage of malnourished youth • Number and percentage of youth living in shelter without basic infrastructure
Level of Education	• Youth literacy rates • Net enrollment in primary education • Net enrollment in secondary education • Gross enrollment in tertiary education • Ratio of female to male enrollment in different levels of education • Educational attainment at age x
Employment Status	• Percentage of youth employed, unemployed, and underemployed • Youth labor force participation rate • Ratio of youth to adult unemployment • Number and percentage of youth not in school, training, or employment • Youth employment by sector (or by type of occupation) • Youth employment by type of work (wage, self-employment, employer, family and unpaid work) • Youth employment by type of contract (formal versus informal, full-time versus part-time, open-ended versus temporary) • Average monthly income • Average time to find a job • Availability of internships and apprenticeships • Young people's job preferences (government, private sector, self-employment, etc.) and attitudes toward work
Health	• Prevalence of HIV/AIDS among youth (in %) • Number and percentage of youth with correct knowledge of HIV/AIDS and contraception • Number and percentage of youth who currently use contraceptive method • Number and percentage of youth who report higher risk sex in last year • Prevalence of drug use (e.g., cannabis, cocaine, inhalants) • Prevalence of youth tobacco and alcohol consumption
Family Situation	• Median age of sexual initiation • Number and percentage of women age 15–19 who have children or are pregnant • Share of female population with at least one child, by age • Ratio of girls married under the age of 18 • Median age of first marriage, by gender
Criminal History	• Number and percentage of youth as perpetrators or victims of crime and violence • Number and percentage of youth members in gangs or rebel groups • Number and percentage of youth affected by domestic violence • Number and percentage of youth affected by sexual violence • Number and percentage of youth in contact with the justice system

Sample Indicators for Youth Assessments (cont'd)

Citizenship	• Number and percentage of voting age youth who vote in elections • Number and percentage of youth who engage/volunteer in community activities • Number and percentage of youth who participate in youth organizations, councils, etc. • Reasons youth do not participate in civic activities
Vulnerabilities	• Number and percentage of youth living in the street • Number and percentage of youth affected by prostitution and/or human trafficking • Number and percentage of conflict affected youth • Number and percentage of youth with mental and/or physical disabilities • What vulnerabilities are specific to young men and women?
Perceptions and Aspirations	• Number and percentage of youth satisfied with their current lives • Number and percentage of youth confident about the future • Number and percentage of youth who want to migrate • Number and percentage of youth who have trust in adults, community leaders, politicians, institutions, etc. For those who not satisfied, what would they like to tell them? • What aspirations do youth have for their education? How do they view and compare vocational education versus university and why? • What do youth express as their priorities, ambitions, and opportunities? • What are the most pressing challenges, risks, and frustrations facing youth?
Direct Environment	
Category	**Selected Indicators**
Family	• Average size of household • Percentage of single parent households • Working patterns of parents (employment status, hours of work, etc.) • Average number of hours parents spend with their children per week • Parent–children relations
School	• Number and percentage youth with a secondary school within x kilometers • Types of schools and training institutes in the area • Quality of existing schools and training institutes in the area • Prevalence of school fees • Teacher perceptions about youth
Neighborhood	• Socioeconomic status and inequalities • Social dynamics within the community (e.g., trust, social networks, etc.) • Prevalence of crime and violence • Peer dynamics
Youth-Friendly Services	• Number of health, culture, sport, religious-based services (public and private) • Share of youth satisfied with services • Reasons for satisfaction or dissatisfaction

Sample Indicators for Youth Assessments (cont'd)

Local Environment	
Category	**Selected Indicators**
Local Economy	• Share of existing firms by sector/industry • Average monthly wages by industry (in $) • Labor demand by industry • Employer perceptions about youth • Skills gaps reported by employers
Local Government	• Budget allocations by sector (infrastructure, education, justice, specific youth programs, etc.) • Existing public initiatives targeting youth
Access and Use of Technology and Media	• Percentage of youth using cell phones • Percentage of youth with access to computers and Internet • User penetration of TV, Radio, etc. • How are technology and media being used? What are they used for?
Access and Use of Financial Services	• Number and percentage of youth with a bank account • Available types of saving • Sources of credit • Conditions to borrow money at local bank (minimum age, collateral, etc.)
Societal Environment	
Category	**Selected Indicators**
Country Demographics	• Population by age and gender • Percentage of population under the age of 25/29 • Percentage of population between the ages of 12/15 and 24/29 • Ratio of male to female population • Percentage of urban and rural population • Percentage of youth population by religion and ethnicity • Percentage of youth population by primary language
Legal Framework	• Age of maturity • Voting age • Legal minimum age of marriage • Legal protections and obligations by age • Customary law
Dominant Beliefs and Ideology	• Social organization and hierarchies • Family lineage • Marriage patterns • Inheritance patterns • Value system

Note: To yield more differentiated information, many of indicators listed above can be disaggregated by age, gender, education level, household income, etc.

APPENDIX 2. Cost Solutions

A little creativity often goes a long way in reducing the costs of the impact study. However, this may also involve a tradeoff with the quality of the evaluation, so each aspect needs to be considered with care.

Reducing the cost of consultants. It is not always necessary to hire an international expert to lead the evaluation. Working with a specialist through a partner organization or university (where the consultant may use the results of the evaluation for a research publication) may save a lot of money. Also, master or PhD students in economics, public health, or other social sciences are often equipped with the necessary analytical background to support the evaluation and survey design and carry out the data analysis and would be happy to work on real-life programs.

Reducing the cost of data collection. Whether and to what extent the cost for data collection can be reduced will have to be determined in collaboration with the lead evaluation expert when the evaluation plan and methodology is being finalized.

- **Piggybacking on existing data.** It may not be necessary to collect our own data if it is possible to build on local, regional, or national surveys that may have been conducted in the past or that are in the planning stages. If good survey data exists, the impact evaluation can be reduced to analyzing this data. Conversely, if another survey is already planned that will cover our population of interest, the evaluation may be able to add a series of questions relevant to our program.
- **Reducing the sample size.** Some programs are tempted to decrease the number of people to be surveyed to save money. While this decreases costs, it also increases the likelihood that the evaluation mistakenly finds no effect of the program when there is in fact one. Using smaller sample sizes must be done cautiously. The evaluation expert hired to oversee the impact evaluation will be able to recommend a minimum sample size that can be used in the specific context.
- **Reducing the length of the survey instrument.** Survey questionnaires tend to grow in length as different stakeholders suggest additional items that it would be interesting to include. By limiting the questionnaire to the major outcomes of interest and leaving out question that are not directly related to the core objectives of the intervention, time and money can be saved during the data collection and analysis.
- **Using existing instruments.** It is not always necessary to develop an entirely new survey for every evaluation. If an organization has a standard program design across different interventions, it is possible to use existing questionnaires from within or outside the organization (or at least parts thereof) that have already been validated. It is often possible to build a set of questions and survey modules that can be used in several evaluations that share the same intervention logic and objectives.
- **Using the program registration process to collect baseline data.** It is sometimes possible to use the natural program registration process to collect data on those who are signing up for the program. This information can include basic sociodemographic characteristics as well as information on outcomes of interest, which in turn, can serve as a baseline.
- **Collecting data with program or local staff.** Collecting data internally instead of

hiring a survey firm can potentially save a lot of money. The same is true for using teachers (when the program is based at a school or training center), university students, or other people in the local community who are willing to work for a rate significantly below a survey firm. This can work well for short and simple surveys but has some important drawbacks, especially for more extensive and larger data collections. Given the complexity of data collection, program staff or other locals usually do not have the experience and skills needed. Supervision and training costs may increase, as well as the time required to complete data collection. If anything goes wrong in the data collection phase, it may mean that the data are useless, and so data collection must start again. This can lead to very expensive mistakes. Also, collecting data with program staff may lead to concerns regarding neutrality and thus the reliability of results.

- **Using innovative data-collection tools.** With new technologies emerging constantly, there is an increasing array of new techniques that can be used for data collection. For example, instead of using paper-based surveys, one can consider cellphone- or computer-based data entry on the spot, thus reducing to literally zero the time and cost for later data entry and processing.
- **Using self-administered questionnaires or phone-based interviews.** Even though it would potentially reduce costs significantly, relying on self-administered instead of interviewer-led questionnaires is usually not a viable data collection strategy for youth livelihood programming in developing countries, especially considering low levels of education. They should therefore only be used in very specific circumstances and after intensive consultation with the evaluation expert.
- **Focusing on geographic areas that can be easily reached.** Costs can be saved by limiting data collection to geographic areas that can be easily reached. If a program is implemented in different sites across the country, the evaluation may focus on the capital city rather than on remote rural areas. However, this will also mean that the findings are only representative for the specific subpopulation and geographic area (such as an urban setting) included in the evaluation.
- **Reducing the baseline survey.** It is not always necessary to conduct a large baseline survey, or, in some cases, any baseline survey at all. For example, when a program is well developed and has a strong monitoring framework, a small baseline survey that collects fast data on participants will be enough to understand their current position and be comparable across the two groups. Similarly, for randomized evaluations, if an evaluation makes use of very large sample size, there is little reason to be concerned about differences between treatment and comparison group, and so a baseline may not be absolutely necessary.

  However, eliminating a baseline entirely is risky. An initial survey means an initial contact with respondents, which often helps to confirm the population of people with whom the coordinators should follow up. It also ensures location information is very accurate for finding people for follow-up surveys. Youth tend to migrate a lot, so eliminating a baseline could greatly increase the cost of finding these people in one, two, or three years. Without a baseline, it is also impossible to use the full list of individual characteristics typically employed for the data analysis. Gender can still be used, but it is not possible to know the well-being of people at baseline.
  Seeking external funding. Financing for an impact evaluation does not have to

be fully covered out of a program's budget. In fact, funding is often complemented with external sources, such as from government, the United Nations, development banks, foundations, philanthropists, or research and evaluation organizations. In fact, the growing emphasis on evidence-based programming and policymaking has increased the availability of funding from a variety of sources. When an evaluation has the potential to fill a substantial research gap that is of interest to the development community at large, as is the case for most impact evaluations in the youth livelihood field, then practitioners can be confident that outside funding may be available.

Postponing follow-up data collection. In some cases, all internal and external funds together may not be sufficient to budget for the entire evaluation with potentially several rounds of data collection. In that case, program staff and the evaluator may decide to limit the initial engagement to a robust evaluation design and the baseline survey and analysis. As funding becomes available in the future, follow-up data collection and analysis can then be added. In practice, accessing incremental funding to pay for follow-up data collection is often easier than trying to get external funding for the entire evaluation.

APPENDIX 3. Verification & Falsification Tests

Randomized Lottery and Phase-in

Randomized assignment methods are the most robust techniques for estimating counterfactuals; they are considered the gold standard of impact evaluation. Some basic tests should still be considered to assess the validity of this evaluation strategy in a given context.

- Are the baseline characteristics balanced? Compare the baseline characteristics of the treatment group and the comparison group.
- Has any noncompliance with the assignment occurred? Check whether all eligible units have received the treatment and that no ineligible units have received the treatment. If noncompliance appears, use the randomized offering method.
- Are the numbers of units in the treatment and comparison groups sufficiently large? If not, we may want to combine randomized assignment with difference in-difference.

Randomized Promotion

Randomized promotion leads to valid estimates of the counterfactual if the promotion campaign substantially increases take-up of the program without directly affecting the outcomes of interest.

- Are the baseline characteristics balanced between those who received the promotion campaign and those who did not? Compare the baseline characteristics of the two groups.
- Does the promotion campaign substantially affect the take-up of the program? It should. To confirm, compare the program take-up rates in the promoted and the non-promoted samples.
- Does the promotion campaign directly affect outcomes? It should not. This cannot usually be directly tested, and so we need to rely on theory and common sense to guide us.

Discontinuity Design

Regression discontinuity design (RDD) requires that the eligibility index be continuous around the cutoff score and that units be comparable in the vicinity of the cutoff score.

- Is the index continuous around the cutoff score at the time of the baseline?
- Has any noncompliance with the cutoff for treatment appeared? Test whether all eligible units and no ineligible units have received the treatment. If we find noncompliance, we will need to combine RDD with more advanced techniques to correct for this "fuzzy discontinuity."

Difference-in-Difference

Difference-in-difference assumes that outcome trends are similar in the comparison and treatment groups before the intervention and that the only factors explaining changes in

outcomes between the two groups are constant over time.

- Would outcomes have moved in tandem in the treatment and comparison groups in the absence of the program? This can be assessed by using several falsification tests, such as the following: (1) Are the outcomes in the treatment and comparison groups moving in tandem before the intervention? If two rounds of data are available before the start of the program, test to see if any difference in trends appears between the two groups. (2) How about fake outcomes that should not be affected by the program? Are they moving in tandem before and after the start of the intervention in the treatment and comparison groups?
- Perform the difference-in-difference analysis using several plausible comparison groups. Do we obtain similar estimates of the impact of the program?
- Perform the difference-in-difference analysis using the chosen treatment and comparison groups and a fake outcome that should not be affected by the program. We should find zero impact of the program on that outcome.
- Perform the difference-in-difference analysis using the chosen outcome variable with two groups that we know were not affected by the program. We should find zero impact of the program.

Matching

Matching relies on the assumption that enrolled and non-enrolled units are similar in terms of any unobserved variables that could affect both the probability of participating in the program and the outcome.

- Is program participation determined by variables that cannot be observed? This cannot be directly tested, so we need to rely on theory and common sense.
- Are the observed characteristics well balanced between matched subgroups? Compare the observed characteristics of each treatment and its matched comparison group of units.
- Can a matched comparison unit be found for each treatment unit? Check whether sufficient common support exists in the distribution of the propensity scores. Small areas of common support indicate that enrolled and non-enrolled persons are very different, and that casts doubt as to whether matching is a credible method.

Source: Reproduced with permission from Gertler et al. (2011, pp. 118–119).

GLOBAL PARTNERSHIP FOR YOUTH EMPLOYMENT

In 2008, with support from the World Bank Development Grant Facility, the International Youth Foundation, the Youth Employment Network, the Arab Urban Development Institute, and the Understanding Children's Work Project joined together to form the Global Partnership for Youth Employment (GPYE). Its goal: to build and disseminate evidence on youth employment outcomes and effective programs to help address the challenges facing young people in their transition to work. The GPYE leverages the technical and regional experience of the five partner organizations in youth employment research, programming, evaluation, and policy dialogue. The partnership's work focuses on Africa and the Middle East, regions in need of better evidence on effective approaches to promote youth employment.

For more information, visit *www.gpye.org*