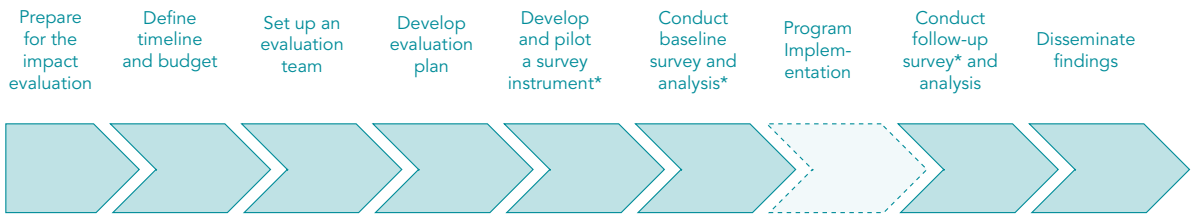
NOTE 7: A Step-By-Step Guide to Impact Evaluation

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

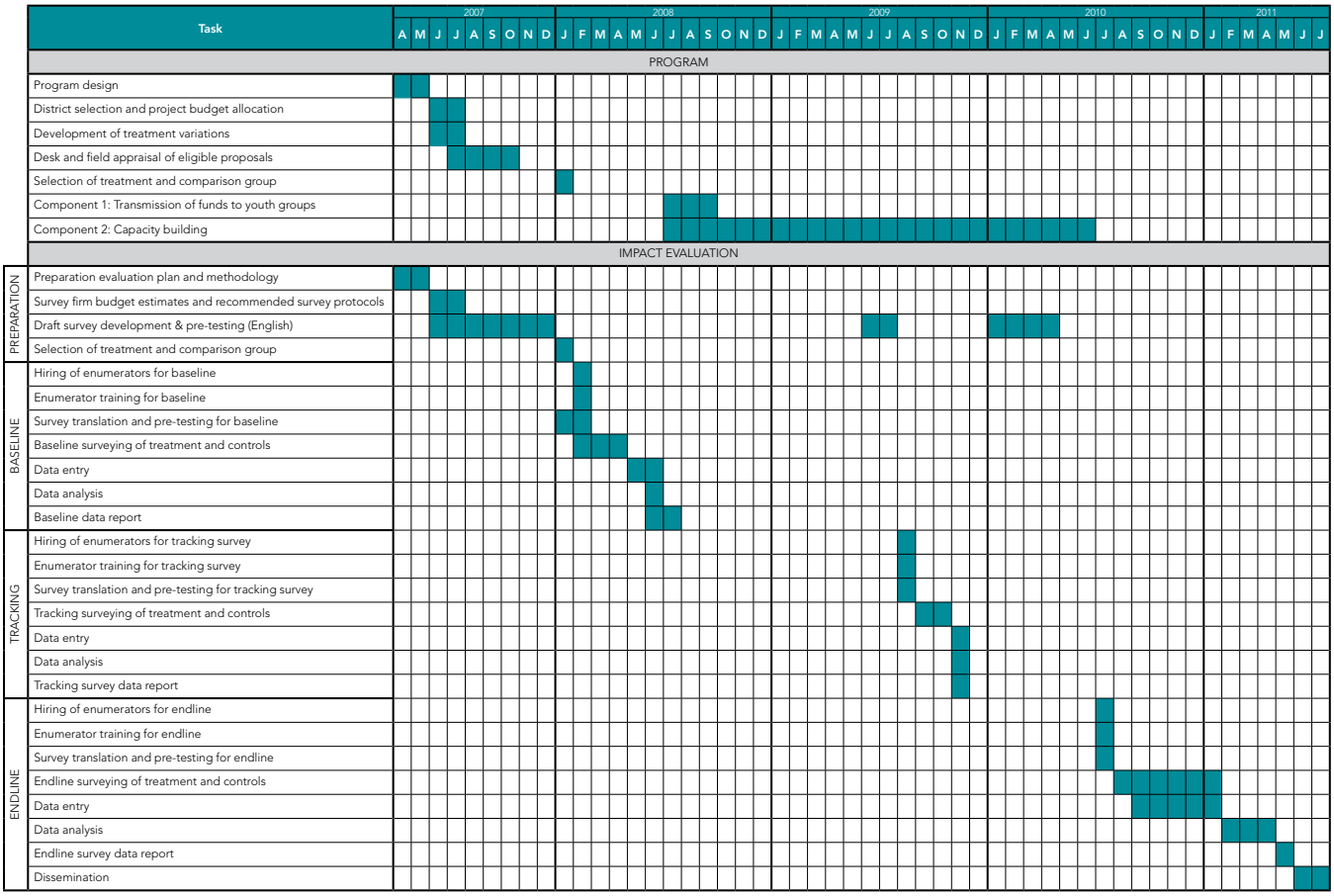
Limited use of the evaluation findings. If the results of the evaluation are not sufficiently shared with internal and external stakeholders, then the evaluation's main objectives of learning for the program and the youth livelihood sector at large are compromised. One way to overcome this issue is to 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