Impact evaluation

A practical guide to designing and administering impact evaluations of PIN programs

Vojtěch Bartoš

Ian Levely

About the guide

This guide presents the basics of **impact evaluation**, including why and when a project should be evaluated and how to design a study that is able to truly measure change, which can be attributed to the project. The results of impact evaluations can be used to select the most efficient project in reaching a particular goal by using a **cost-benefit analysis** approach. After reading this guide, you should be able to make informed decision about how to **evaluate a project** and **learn from its successes or failures**.

About the authors

Vojtěch Bartoš , CERGE-EI, Prague, bartos.vojta@gmail.com.

Ian Levely Institute of Economic Studies at the Charles University, Prague, ianlevely@gmail.com.

Both authors are members of the Group for Analysis of Poverty and Inequality (GAPI). www.gapiresearch.org

Phone:(+420) 226 200 400
Email:mail@clovekvtisni.cz
Web:www.peopleinneed.cz

Table of contents

Chapter 1: Introduction to evaluation methods

- Evaluation is not only for donors, but mainly for the future use of the organization for improving quality of services provided and deciding which methods are most effective (cost-benefit analysis).
- With limited resources to spend on program evaluation, prioritize components of the program that are applicable to future projects, for scaling up, or in other settings.
- If two approaches to address similar issues are used, evaluations can determine which one is most cost effective.
- Looking at selected indicators should give you an answer as to whether the living conditions of the household have indeed improved, if this is your goal.
- Evaluation should also give you a good reason for why the program works, so that you are able to replicate it even in different conditions.

Why evaluate?

The most immediate answer to this question is that evaluations of some form are usually required by donors. While this isn't a perfect incentive, this chapter offers some ways in which properly designed impact evaluations may actually help you better understand which parts of a program work and, if there are multiple approaches to achieving a similar goal, you may learn which of the approaches delivers most good for the least money, i.e. which is most cost-effective.

Conducting an evaluation may be costly, as it requires collecting lots of data both before and after—and sometimes during—the project. Measuring long-term effects is desirable, as sustainability should the ultimate goal of every development project.

Monitoring vs. Evaluations vs. Needs assessment

First, it is important to make a clear distinction between three types of research related to development projects: monitoring, evaluation and needs assessment. While this part of the guide deals primarily with evaluation, Part 3, which is devoted to data collection, applies to all three types of studies. The table 1.1 presents the key features of each.

Table 1.1: Differences between monitoring, evaluation, and needs assessment

	Needs assessment	Monitoring	Impact evaluation
What	Measuring characteristics of target population	Measuring intermediate outcomes (Are people receiving and using the intended assistance?)	Measuring the effect of the intervention on individuals' wellbeing (i.e. health, wealth, education, awareness, etc.)

Why	Identify potential beneficiaries and properly tailor the program to suit their needs	As part of impact evaluation: intermediate targets have been met. / Identify	Determine whether the program works, why it works (or does not) and if so, how large is the effect. Use the results for further project scaling-up or for cost-benefit analysis.
Who	Random sample of the target population	Beneficiaries / project staff / other involved parties	Treatment (beneficiaries) / Control or comparison group
When	Before designing the intervention	During the intervention	Both before and after the intervention

Source: Authors

A **needs assessment** is done before the project begins and the results are used to identify what interventions are appropriate in the geographic area of interest. As a needs assessment is conducted before the project has been designed, it is fundamentally different from monitoring and evaluation, both of which are studies that measure the effects of existing projects.

Monitoring consists of measuring *outputs* provided by the NGO. In other words, making sure that the project is running smoothly and that resources are finding their way to beneficiaries as intended.

An **evaluation**, on the other hand, measures *outcomes* of the projects: the ultimate change in beneficiaries' quality of life (income, health etc…). We can further divide outcomes into short-term outcomes, which are apparent soon after the project concludes, and final outcomes, which are those effects of the program that are self-sustaining in the long-term.

Results chain

The results chain table below, provides a framework for thinking about how activities undertaken by an NGO (or development agency) convert project inputs to outputs, and subsequently how those outputs lead to outcomes over the short-term and final outcomes over the long term. This is a useful way of organizing thoughts and identifying precisely how and why the project is expected to produce the desired results and to identify any shortcomings in the design.

Table 1.2: Results chain

Results chain				
Supply side (mostly under the control of the NGO)			**Demand side (not under NGO control)**	
Inputs	**Activities**	**Outputs**	**Outcomes**	**Final outcomes**

Resources at the disposal of the project -- Budget -- Staff Physical resources -- Cars / motorbikes -- Office space Local counterparts Know-how	Conversion of inputs into outputs -- Developing training materials -- Training of staff or third-parties involved in the project (e.g. teachers) -- Procuring resources for the project	Tangible goods and services delivered as a part of the intervention -- Trainings based on the developed training materials -- Distributions (cash or in-kind) -- Establishment of necessary infrastructure -- Other services provided to clients	Results likely to be achieved if the beneficiaries use the project outputs -- Lessons from the training implemented in practical use by the beneficiaries -- Increase in well-being due to training or provision of goods	Results sustainable over the long-term -- Long-term change in well-being and behavior of the beneficiaries
Example for FFSs				
-- Staff (and staff of partner NGOs) -- Existing knowledge of agricultural methods and training materials for FFS. -- Physical resources owned by PIN and partner NGOs (vehicles, farming equipment etc.…)	-- Development of new training materials, including research into which methods are appropriate for the given area. -- Training instructors -- Acquiring seeds and tools for distribution -- Implementing FFSs	-- Farmers trained in improved agricultural techniques -- Model fields established in communities -- Seeds and tools distributed to farmers	-- Direct beneficiaries use improved techniques on their own fields -- Use of high-yield seeds by farmers -- Improved yields for beneficiaries (in current season)	-- Long term adoption of improved techniques in further planting seasons -- Indirect effects: neighbors of beneficiaries use improved techniques as well -- Improved long-term welfare in targeted communities

Source: Authors

What to evaluate?

A natural window to begin an evaluation is during the pilot phase before the program is scaled up. This is the opportunity for the implementing agency to rigorously assess and test the effectiveness of the program and to improve its design. It is the high cost that should make you think which activity to evaluate in the first place. Evaluating each component may be interesting, but the benefit one would achieve is unlikely to equal to costs spent on conducting the evaluation.

Assessing which project components are worth evaluating is somewhat subjective. You should consider the benefits of knowing the effect of the particular component on the indicators of your choosing. Typically, a component is worth evaluating if you plan to conduct a similar project in the future, or if you want to scale the current project up but are unsure if it has been successful. Another possibility is that there may be multiple possible approaches to the same goal and you want to test which approach works best.

The reason for testing a particular project component, rather than the project as whole, is that the primary aim of each evaluation should not only be to understand *if* the project works, but also *how* the project works. Taking the example of the results chain for farmer-field schools from table 1.2, you should ask whether the effects you observe are caused by the distribution of seeds or by the seminars the farmer received alone, or whether both of these interventions necessary as complements. If the results are driven by the seminar alone, then the benefits of the program could be increased substantially by concentrating on this component of the program.

Selection of program indicators

The selection of the *indicators* is crucial as well. The indicators should not create a checklist of what has been completed in the project. Rather, indicators are a list of quantitative variables that measure the desired effect of the project. The variables measuring the completion of project, such as number of farmers being delivered an improved wheat seed, or the number of bags of wheat the farmer collected. These are the means by which we achieve the variables of our interest: did the household increase its consumption? Has the health status of the household improved? Have the children started visiting school more often or they dropped out of school less?

A well thought-through results chain can provide you with a good understanding of which indicators should be selected and when it's a good time for their measurement. You should obviously think of both indicators related to program monitoring and to program evaluation, but it is important to understand the difference. You should consult the results chain table above.

In order to select proper indicators, the pneumonic acronym SMART is a good guideline:

- **Specific:** information as detailed as possible
- **Measurable:** information can be acquired
- **Attributable:** indicator is linked to the project's goals
- **Realistic:** data can be obtained in a reasonable time, and at reasonable cost
- **Targeted:** covers the population of your interest

Imagine that you want to evaluate the effect of an intervention that distributes wheat seeds to farmers by examining statistics that measure the amount of crops harvested by farmers in the program, but you measure only the harvest of wheat. In this case, you would not observe if, for example, households switched to wheat entirely and stopped planting corn, bringing the household the same level of income as before, only reducing the variety of crops planted, hence increasing risk of crop failure. Or even if you can observe this effect, but you do not measure expenditures on the improved seeds you cannot assess if the household is actually better off at the end, which is your primary goal.

On the other hand, if you measure the household welfare indicators and you compare these to a proper comparison group that did not participate in the project, you can immediately see if the household improved its wellbeing and how. We will talk about how to create such *comparison group* in the next section.

Indicators for some projects may also be subtler, such as involvement in communal activities or perception of women in politics. Even such variables can be quantified by asking about ranking certain opinions on a scale, for example. We should also ask if the effect of the project is lasting or if it is only due to expectations of future benefits that evaporate when the project ends.

External validity

Last, you should also be aware of limited scope of your evaluation. Unless you have a very good understanding why the particular project helps the way it does, you will not be able to know what effect would a similar project have in a different environment, i.e. in a different country or in a different social group. This is known as the problem of *external validity*.

To conclude this section, once you start considering an evaluation of a project:

- You should know why and what to evaluate.

And once you know this and when you are designing the evaluation, you should:

- Be aware of how the information you learn from the evaluation helps you.
- Know what questions, i.e. what indicators help you give this information.
- Try to understand what was the reason for the effects you observe.

Who to evaluate?

In the previous section we discussed when and why to do an evaluation. In this and the upcoming sections you should learn what it takes to conduct a proper evaluation.

Usually, we are interested in learning the effect that a program implement has on wellbeing of the beneficiary. However, comparing the same individual over time will not, in most cases, give a reliable estimate of the program's impact. Many other things affecting the outcomes may have changed since the program was introduced. Thus, we cannot get a proper estimate of the impact of the program on a given individual. What we can do is to obtain the average impact of a program on a group of individuals by comparing them to a similar group of individuals who were not exposed to the program. As such, impact evaluations can be compared to a clinical drug trial in which one patient receives a treatment and the other does not. By comparing our outcomes of interest, for large enough groups of those receiving and not receiving the treatment—those who participate in the program—we can learn what the true effect of the treatment on the targeted group is.

Know what would have happened if program was not implemented. These groups are called a *treatment* and a *control group*. It is obvious that in order to be able to draw reasonable conclusions, we need to make sure that the groups are as similar as possible before we actually start implementing the program. Ideally, we would want the two groups to be exactly the same, i.e. to create a "parallel universe" where the program is implemented in one and not in the other. The next section will discuss the gold standard of evaluation: randomized control trials (RCTs), in which individuals are randomly assigned into the program. Later, we will also discuss other possible ways of creating a reasonable comparison group.

Know your population of interest. For now, we shall discuss *who* you should evaluate. This question requires careful consideration, as the purpose of the evaluation may be different from case to case. Sometimes you are interested in learning the effect of the program on the population you are targeting. Sometimes, you want to know if the effects of the program spill over to other members of the community. Sometimes, you may be interested in studying the effect of the program on a particular population or subgroup. In each of case, both the treatment and control group would be created differently. You need to decide in advance, which of the approach is appropriate and adjust the evaluation design accordingly.

Example 1: Spillover effects

As a first example, imagine that you evaluate the effect of a seminar on improved agricultural techniques. Here you should take into account that the information is likely to spread across the village. Hence, if you select the control group from the village in which you give the seminar, you are most likely to end up with an underestimation of the actual impact, as the fellow farmers from the village also learn, indirectly, and apply some of the methods on their own fields. Benefits to those who have not directly participated in the program are called *spillover effects.* When assessing the costs and benefits of this approach, such a design would lead you to omit gains made by those indirect beneficiaries as a result of the intervention. This is a double loss, since you would not only erroneously conclude that the program wasn't working, but you would do so because you would be ignoring positive effects on the entire community. Thus, for such a program it is better to select the control group in another village. In the next section we will discuss the selection of beneficiaries using randomization on different levels (individual or village) and you will see that this would be a perfect candidate for randomization on a village level.

Example 2: Imperfect compliance

A second problem arises because the seminars are not obligatory and some of the individuals to whom the program is offered might chose not to participate. Here, you are interested in multiple outcomes: firstly, the effect of the program on those who actually *participated* in the course, and secondly, the effect on everyone to whom the course was offered, taking into account those who did not participate. Lastly, you would want to know what types of individuals actually participated and why. This is why you should collect the data on both the participants as well as on those who did not participate, including those who drop out, in order to learn about their motivations and any obstacles that prevent people from taking part.

More generally, when doing an evaluation, we are interested in two principle measures: the *Average treatment effect on the treated* (ATT), which is the average effect of the program on individuals who actually took part in the program and the *Intention to treat* (ITT) effect, which is the average effect of the program on all individuals to whom the program was available, regardless of whether they participated or not.

Table 1.3: Different treatment effects

Measure of interest	Treatment Effect on the Treated	Intention to Treat	Spillover effects
Effect measured	Effect of program on those who directly took part	Average effect of program on all individuals to whom the program was available	Effect of the program on those from communities where the program was implemented
Treatment sample	Participants	All those eligible for the program	All community members
Ideal control sample	Individuals who meet criteria and were willing to participate, but to whom the program was not made available.	People who meet criteria for participation, but to whom the program was not made available.	Members of similar communities where the program was not implemented.
When to use	If you are interested in learning the effect of the program on the participants.	If you are interested in learning the cost effectiveness of your program (even if not everybody participates).	When you expect wider impact of the program.

Source: Authors

Intention to treat vs. Average treatment effect in a deworming program

One of the success stories of development interventions is a deworming program. It has been shown that deworming is actually the most cost-effective way how to increase school attendance. Imagine you implement a deworming program as a part of your project. You want to treat every child in every class, but not everyone is in school on the day the medication is administered. If you are interested in the *average treatment effect* on the treated, you actually want to know the true effect of the medication on a particular individual who was treated. To learn the effect you compare the data only for the kids who actually took the deworming pills with similar kids in other classes who were not offered the program

On the other hand, if you were interested in the *intention-to-treat effect*, you would study the average effect of the program including those who didn't actually participate. This allows you to measure the cost and benefits for the individuals who drop out or don't participate for their own or other reasons. To learn this type of effect you compare the data for every kid in every class that participated in the program with similar classes that were not offered the program.

Key reading

If you need more detailed information on impact evaluations, you might want to refer to the World Bank's Impact Evaluation Toolkit: Vermeersch, Rothenbühler, Sturdy (2012): Impact Evaluation Toolkit: Measuring the Impact of Results-Based Financing on Maternal and Child Health. World Bank. Available online at: http://go.worldbank.org/IT69C5OGL0

Another well-designed impact evaluation manual is the following: Hampel, Fiala (2012). Measuring Success of Youth Livelihood Interventions: A Practical Guide to Monitoring and Evaluation.Washington, DC: Global Partnership for Youth Employment. Available online at: http://www.gpye.org/measuring-success-youth-livelihood-interventions

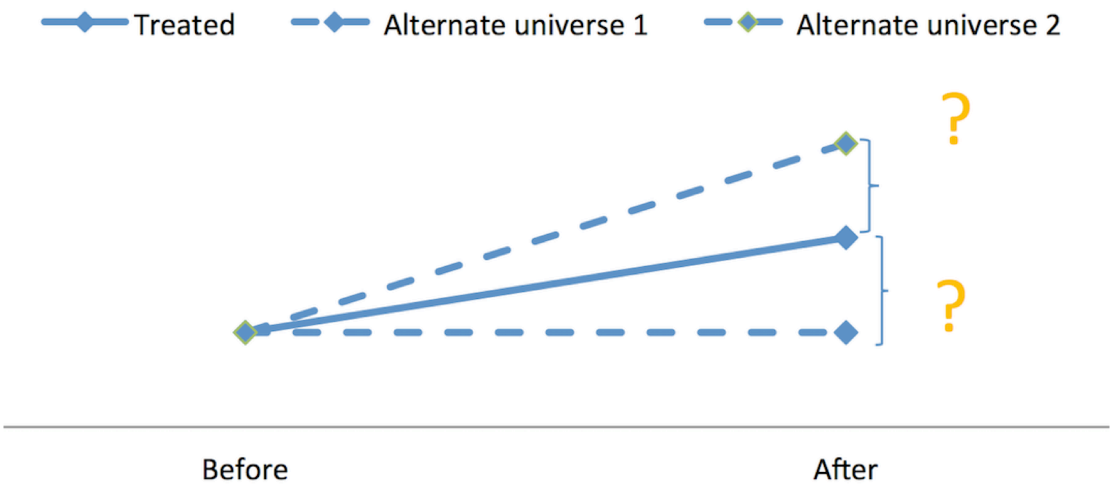
Chapter 2: Randomized control trials

- Well-designed randomized control trials (RCTs) are the golden standard of evaluation, which allows for exact measure of the effect of the program.
- Surveying the people randomly selected into a control group can tell you "what would have happened in the absence of the program," which is necessary for assessing the real impact.
- Impact evaluation is a first step on the long way towards finding the most effective approach that is most beneficial for the people. Without this knowledge you cannot be sure if your program helps, has no impact, or worse, harms.
- This chapter also discusses issues of sample size, problems of attrition, methods of randomization and discusses how to look at effect at heterogeneous parts of the targeted population.

In the previous chapter, we discussed the necessity of having a proper control group for our evaluation of the effect of the program on the treated group. Imagine selecting half of the poorest villagers in a village as your treatment group and enroll them into the seminar on agricultural techniques and for comparison, you would pick the remaining, richer half. This can be problematic, as we may easily see that we are comparing apples and oranges. This example would obviously lead to underestimation of our project. The richer households simply had more in the beginning and they would be, say, more resistant to natural disasters.

The problem in the following example is less obvious: we select a group of participants for our agricultural seminar based on pre-selected criteria such as a poverty index and land ownership. Half of the eligible villagers finally decide to take part in the seminar, so this is our treatment group and we decide to track the remaining eligible villagers as a control group. These groups may be, on average, initially comparable based on all the criteria we measured. However, there may be important *unobservable* differences between the groups that are not possible to measure. For example, the simple fact that one group decided to participate and the other did not suggests an important difference, which may actually be substantial, as the participating group may be more pro-active. This would lead to actual overestimation of the effect of the program. So how do we create a proper control group?

Figure 6.1: Measuring impact with a counterfactual, "parallel universe"



Source: Authors

Creating a proper control group

An ideal way to create a proper control group is to select a group of potential program beneficiaries and randomly select half of these people to participant, then to track both groups for comparison. If the program does not allow for explicit random selection, you can still look for some *random* pattern: Are there similar individuals in similar communities where the program was not available? You can also try to exploit the random timing of the program implementation if the program was staggered in some areas. The following lines will try to convince you that randomization is not only often feasible, but that it is also a fair process of beneficiary selection.

Why we need a proper control group and why randomization solves the selection problem?

Imagine you run a farmer field school program and only collect data from program participants before and after the project. You find that consumption increased dramatically for these individuals during this period. Can we attribute this to the program? It would seem desirable, but in reality we cannot be sure. There are many possible explanations, which you will not be able to distinguish, unless you have a proper comparison group, your *parallel universe*. What if the harvest in the baseline year was particularly low, thus *everyone's* consumption has risen during the period of implementation—regardless of participation in the project? Unless you have a proper control group, your data will not tell you what part of the increase is due to the project and what is due to unrelated factors such as weather.

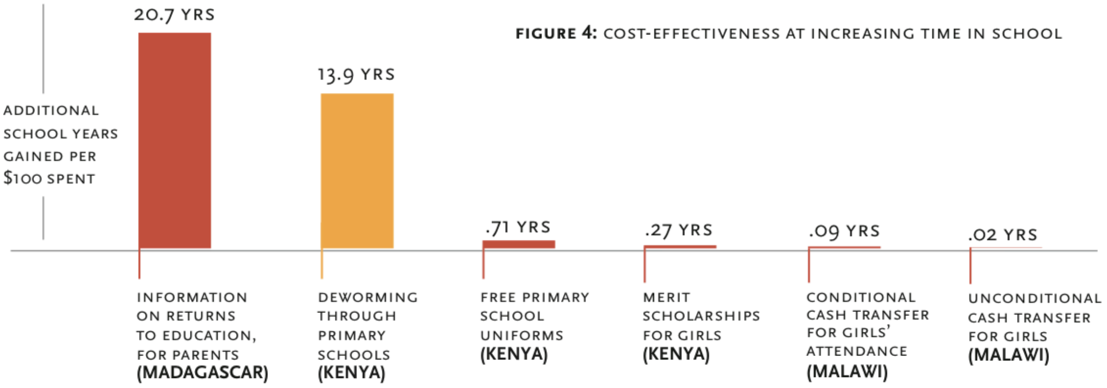
Ethical issues of randomization

The random selection process is sometimes seen as unethical, as some people will invariably be denied services based chance.

Uncertainty of impact of the project. Firstly, we should take into account the purpose of the evaluation: to learn if the program works, if it is the most cost-effective, and to decide if to scale the program up or to replicate it elsewhere. Thus, once we have the best knowledge about the actual impact of the evaluated program and possibly also a comparison of multiple possible approaches, future beneficiaries may already be targeted with the best program at hand. Note that before conducting an evaluation we cannot be absolutely sure if the program has any impact in the first place or even if there are no unexpected negative (side) effects.

The Figure below shows what lessons we can learn from properly conducted evaluations of multiple approaches aiming at increasing time in school. We already discussed that an unlikely candidate, a deworming program, may be the most cost-effective way of increasing time spent in school by children at (almost) the lowest cost. Without conducting quality impact evaluations, this result would unlikely ever come up.

Figure 2.2: Cost-effectiveness at increasing time in school



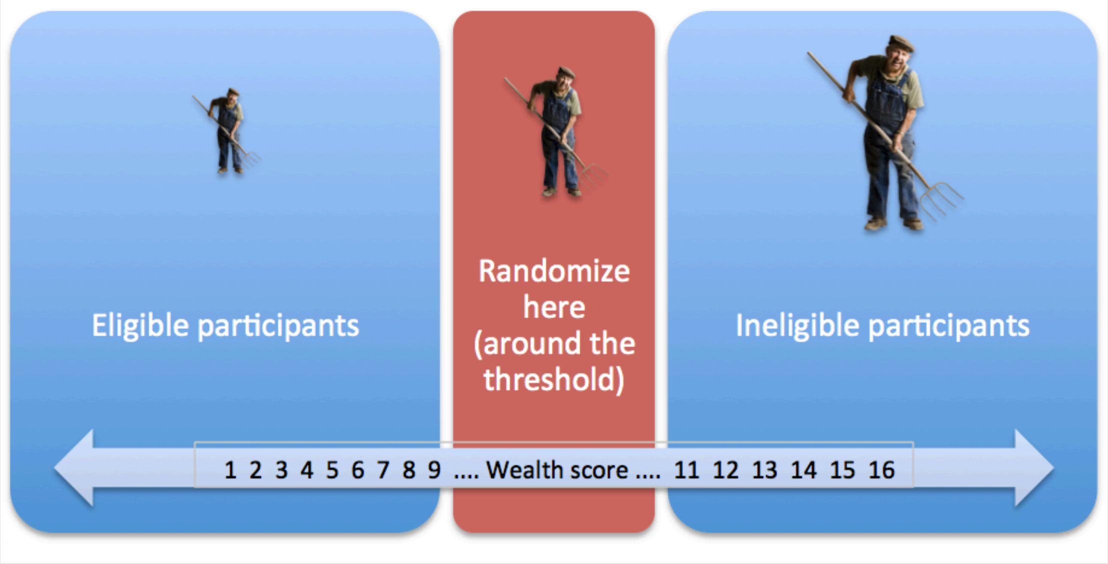
Source: JPAL (2011)

Limited budget and excess demand for program. Even if this is not convincing for you, think of the usual beneficiaries' selection process. You are always limited by the available project budget, thus there is selection going on anyway. There is no general methodology that would help you select the most suitable beneficiary for any particular program. There is always a larger group of potential beneficiaries satisfying the preselected conditions and the final selection is usually *ad hoc* or based on some arbitrarily constructed index. A simple lottery that satisfies the randomness we want to achieve is actually equally as fair if not fairer.

Exploiting a tentative cutoff for randomization

An interesting approach of how to carry out random assignment is the *randomized discontinuity design.* As said, often we have to make ad hoc decisions about the cutoff levels for eligibility in the program. We can exploit this uncertainty and introduce fairness into this selection game by giving a chance to a slightly wider range of almost equally eligible population and using only those on the margin of being eligible or not for our evaluation purposes. Of course, the estimation of effect would be valid for this specific subgroup *on the margin*, but we may have some assumptions about the effect on those further away from the threshold. We will talk more about alternative ways of exploiting discontinuity in the next section.

Figure 2.3: Randomizing on a tentative cutoff


Eligible participants
Randomize here (around the threshold)
Ineligible participants
1 2 3 4 5 6 7 8 9 .... Wealth score 11 12 13 14 15 16

Source: Authors

Moreover, the control group does not necessarily be left without any program participation. Usually, the project is a bundle consisting of multiple programs. For the evaluation you want to end up with a group that takes part in this program and another group that does not take part in the program, but is otherwise equal. Thus we can easily imagine an evaluation design in which one group takes part in one program only, while the other takes part also in another. The only condition is that the program outside our evaluation is not dependent on the implementation of the evaluated program. If this were the case, we would end up with an overestimated effect of the program at the end.

Another possibility is so called *phase-in design*. This approach gives a chance to participate for everyone, only one group is invited to participate earlier and the control group is promised to participate for example a year after, if we think that the project may show some effect already one year from being implemented. When beginning the intervention, we randomize the order of who receives the treatment. While all individuals will eventually receive the same services, the order of program participation is randomized by lottery. We need to think carefully, however, if there is not a chance of the other group to change its behavior in expectation of future enrollment. This would make that group different and it would cease to be a control group we are looking for.

Expectations of being enrolled later in a phase-in design

Imagine that you run a microfinance program in which loans are granted to participants randomly. Those who are not chosen to get loans now will get them in the next phase of the project. If individuals who don't get loans are aware of this, they will likely delay investment until they are eligible for the program. This will make the program seem more effective than it actually is. Thus, this project is not a good candidate for phased randomization.

Methods of randomization

In the previous sections we have discussed why random assignment to program helps us to be able to carry out a proper impact evaluation and we tried to convince you as a reader that ethical issues with randomizing if people are selected into programs is not unethical as it may seem to some. In this section we will look at some practical issues we would often be dealing with when carrying out a randomized impact evaluation: Optimal sample size, level of randomization, attrition and estimation of the impact on various groups.

Optimal sample size

The assumption that the two groups, the control and the treated, are exactly the same relies heavily on statistics. If you imagine an extreme of selecting only two people randomly picked from a targeted population and then randomly assigning one into the program and the other not, we can be almost sure that these people would be different in many aspects and our inference based on observation of the difference of their post-intervention outcomes would not give us any meaningful results. Larger population is hence required for us to be able to make a reasonable inference about the actual impact of the program evaluated. But how large? Unfortunately, the answer is complicated and usually *power calculations* (a statistical method) are used to estimate the optimal sample size. For this, you would want to call a statistician to assist you. In an appendix to this guide, you can find a simple excel form for estimation of a sample size.[1][2]

Once we expect that the effect of the program is likely to be large or that it would have almost the same effect on everyone participating – i.e. the variance of the effect on participants would be small – only small samples may be sufficient (in extreme, if the effect is the same for everyone, one treatment and one control participants are perfectly sufficient for the evaluation). For minimal size of an effect or its high variance we may require much greater samples, though.

Sometimes we are not interested in measuring the average effect only but we may be interested in studying the effect of the program on particular sub-populations. Note, that if you want to see if the program has different effect on the older and on the younger halves of your participants, you need to increase the sample two-fold. Measurement of heterogeneous effects will be discussed below.

You should also know what your level of randomization should be. Once you measure an effect of the program on individuals, individuals are also your units of measurement. If it is villages, it is villages. Luckily, for villages we expect that the results would have much lower variance of the effect, as so many different people on average live in each village and the effect, also due to the law of large numbers, gets closer to the population average anyway.

When to randomize on individual and when on village level?

In the previous section we discussed whom to evaluate. We have shown that in some cases we are interested in results at the individual or household level. This is in case when we do not expect any spillover effects to be present. Examples of such programs may be in-kind donations or cash for work programs, which are not likely to affect much anyone else except for the direct participant of the project. In such case you can offer the program in all communities, and randomly select individuals within each community.

On the other hand, often you run a program indirectly affecting also the non-participants. When evaluating such programs, a within community randomization could be problematic as the non-participants may easily benefit from the program by direct interaction with the participants. Sometimes, spillover effects are an important feature of the project and it would be a pity not to be able to track such effect. Moreover, the average effect of the program would most likely to be underestimated, because we would not be able to account for the improving conditions of indirectly benefiting non-participants. When this is the case, it is in your best interest to do the random selection of entire communities rather than of individuals. However, randomization on a village level comes at a cost. We need to have a reasonably large sample to be able to make a causal inference.

But there may also other types of randomization, not only on the individual or on the village levels. We can also imagine some educational program that affects only a

[1] Hampel, Fiala (2012) also provide an excellent detailed online guide on power calculations here: http://www.iyfnet.org/sites/default/files/gpye-m&e-resource6.pdf

[2] Available online here: https://www.dropbox.com/s/2wsd0zofxbx3apo/power%20calculation%20tool.xlsx

particular class, within which the information may spread, but there is no effect across classes – here, randomization on a class level is an appropriate solution.

Problem of attrition

We have already discussed what to do when someone who is selected decides not to enroll into the program. A greater problem for our evaluation, even if the participants are selected randomly, occurs when the beneficiaries do not complete the program for reasons having to do with their personal characteristics and non-random factors. This problem is called *attrition* and can lead to biased results.

To learn if the attrition is random or not is also one of the reasons why we collect the quality baseline data. When properly randomizing the beneficiaries into treatment and control, we would not need the baseline data for the final evaluation as the effect is wholly captured in the final survey conducted after the program. However, the baseline data help us first of all to see if we managed the randomization properly. Second, we can also see if the problem of attrition is random or if it is driven by some characteristic that we observe.

Let us get back to the example of the agricultural seminar. Imagine that we observe that only 70% of the beneficiaries actually finish the program. If we look at the baseline data, we can see if there are any characteristics that can tell us who left the program and what was the reason. If we cannot see any pattern, it may be that attrition is random and we do not have to worry. If, however, we find that poor farmers are more likely to drop out, we should be concerned. It may also happen that the seminar is not interesting for some particular group of able farmers. If not accounting for such issues, we may easily end up with overestimated or underestimated program effects, respectively.

In any case, tracking participants *even after they leave the program* can help to account for the problem of attrition that is non-random. If tracking all participants who drop out is too costly, you can *randomize* who is to be tracked from the people who dropped out and you would still end up with the data you need. If we track the people who dropped out, we can better understand the reasoning for their behavior.

Finally, note that we want to include data for these quitting people into our final evaluation. This is important, as we cannot simply forget about this particular group of people. Most likely, they would be present also in the group of beneficiaries who we select for the project once we implement it next time or if we plan to scale the program up. The estimation of the effect of the program would tell you what is the *average treatment effect on the treated.* Next, we will learn how to estimate the effects of the program on particular sub-groups of the targeted population, such as the group that is dropping out.

Heterogeneity effects

As previewed, often we want to understand how the program affects different sub-groups, not only the studied population as whole on average. This is another reason why we need to collect the quality baseline data. These data would allow us to establish the sub-groups, which can be constructed based on any observable characteristic. For example, we can select the poorest half of the control and the treatment groups and see what is the effect for this particular sub-group. The same exercise can be done for the richer half and we can make learn for which group is the program more beneficial. Any *pre-intervention*, i.e. baseline, indicator may be used.

Why may effects on subgroups be of so much interest for us?

There is a nice example from a program providing textbooks to children in rural Kenyan primary schools. Glewwe, Kremer, and Moulin (2004) studied this using an RCT and they found no effect of textbooks on students' achievements on average. When they dug deeper into the data, they discovered that there was a positive effect on initially high scoring students. Why was that? Since the textbooks were in English, they were unlikely to help the weaker student.

Further, one can also select the population based on multiple characteristics. For example we can look at sub-groups by income and age or by gender and education level. The larger our studied sample is, the deeper we can dig into the data. This is also an important factor when we are thinking about how many people to follow for the purposes of the evaluation. We discussed this in one of the previous sections. Obviously, this comes at a cost and we need to be sure that the additional resources we spend on collecting data for more individuals would justify the additional information we learn. Usually, we would not be interested in interaction of more than two variables.

Key reading

For more detailed information on RCTs, refer to Duflo, Glennester, Kremer (2006): Using Randomization in Development Economics Research: A Toolkit. JPAL, MIT. Mimeo. Available online at:

http://www.povertyactionlab.org/sites/default/files/documents/Using%20Randomization%20in%20Development%20Economics.pdf

Note: the toolkit may be technically complicated in some parts, but most of the technical details can be skipped and the remaining information will still allow for designing a reasonable RCT.

Appendix to Chapter 2: Source of funding for impact evaluations

Here is a list of selected organizations that provide funding for impact evaluations:

- **3IE:**http://www.3ieimpact.org/grantsoverview/
- **ILO's Youth Employment Network:** http://www.ilo.org/public/english/employment/yen/whatwedo/projects/evaluation_fund.htm
- **UK Economic and Social Research Council (ESRC):**http://www.esrc.ac.uk/funding-and-guidance/funding-opportunities/international-funding/esrc-dfid/index.aspx
- **USAID:**http://grants.gov/applicants/find_grant _opportunities.jsp
- **World Bank's Strategic Impact Evaluation Fund:**http://go.worldbank.org/YM02GKKFJ0
- **International foundations:** Bill & Melinda Gates Foundation, W.K. Kellogg Foundation, Ford Foundation, Mastercard Foundation, Nike Foundation, Hewlett Foundation, and JP Morgan Chase Foundation.

Chapter 3: Other statistical methods used for impact assessment

- Differences in differences, discontinuity design or matching on observables can be used instead of RCTs if randomization is not possible due to any reason.
- All these methods are inferior to an RCT, but offer you a better estimate of the impact of the program than a simple before/after comparison.
- Difference-in-differences method looks at the differences between two similarly evolving groups and whether and how this difference changes over time.
- Discontinuity design is suitable when an administrative threshold is used for selection. Comparing people just above and just below the threshold is as good as randomization, but results are valid for this small sub-sample only.
- Matching on observables might be used if you have access to a large (say, national) database with a similar set of indicators that are of your interest. You might want to find a statistician who might help you to use this database to create a control group of "similar enough" people.

Sometimes it is not possible to randomize people into treatment and control, but even if this the case, there are other statistical methods in that make evaluation feasible. Randomization might not be possible because a donor or partner does not allow for it, or the program has already begun. As such, some methods we discuss in this chapter may help you even in this case of having no comparison data – you might exploit some secondary data for your evaluation.

In the absence of a randomly created *control group,* the next best thing is a *comparison group* of individuals that do not receive the treatment but are similar in certain ways. Here we discuss a couple statistical methods by which it is possible to identify the effects of a program without randomization. The proper comparison group depends on the method of analysis, though a good baseline survey is important in deciding who to include.

Obviously, the main indicators that we are interested in must be included in the baseline survey too, so that we have a meaningful comparison of before and after the intervention. Any comparison of post-intervention characteristics makes no sense unless we know the initial conditions of each group.

As we will see next, sometimes we may be interested in collecting many more observations on the comparison group than on the actual treatment, as we no longer rely on the benefits of random assignment. We may benefit from matching pairs of program participants and non-participants based on their characteristics. In other words, we may compare the people who are similar to each other. There are two possible simple ways of how to conduct an impact assessment without randomization that we discuss here: difference in differences, and discontinuity design. Two other statistical methods, which require assistance of a statistician, are introduced at the end of this chapter.

Reminder: Why is a before/after comparison misleading?

We have already discussed why a simple before/after comparison may lead to biased estimation of the effect, as so many different factors may be driving the difference. If a flood affects the area in which we deliver an agricultural program, we would simply end up with a conclusion that the program was a complete failure. On the other hand, a program may not have any effect at all but a railway just started operating in the area boosting the trade and improved overall economic situation. Using the before/after comparison, we would attribute this effect to our program entirely.

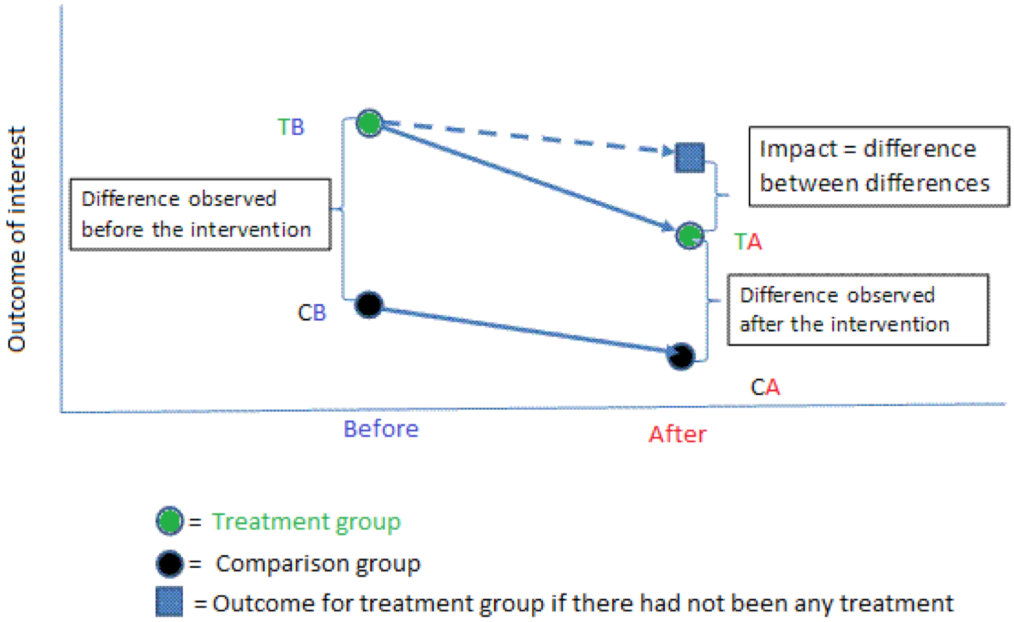
Difference in differences

In Chapter 2 we discussed how to evaluate a farmer field school. Now, assume that we *cannot* randomize the villages randomly, because they have been pre-selected by a partner NGO based on some criteria (possibly these are villages where the NGO has worked before or these villages have applied to take part in the program). We cannot directly compare outcomes from these villages with others in the area since they may differ significantly. The selected villages may be richer (or poorer) than average villages in the area for some specific reasons, and these could be the same reasons that those villages ended up taking part in the program. However, there are still some villages in the area that are located in the same area that is being affected equally by the same shocks, such as droughts or crop diseases. Also, the villages are equally far away from the closest city. If you assume that the villages would react similarly to external shocks—like weather, changes in the market, or diseases—you can use the method of *difference in differences* to assess the causal effect of the farmer field schools on the targeted population.

To examine if this assumption is valid for the particular case, you may look at the baseline and/or historical data for the area. You can also look for anecdotal evidence that no area-specific shocks occurred during the relevant time period (e.g. local flooding, local outbreaks of disease and other). This will give you an idea of who will react to the common shocks in what way and if there are no shocks that would be village specific only. If the participants who enter the program are more sensitive to agricultural shocks, the two groups will react differently, and the comparison will not yield the true effect of the program.

However, if you see that the assumption of similar development trends for both groups is valid in the absence of the program, you can look into the effect of the program and separate it from other confounding factors. You can do this simply by subtracting the difference of the post- and pre-intervention outcomes for the comparison village from the same difference for the treatment group. The figure bellow may be more intuitive in explaining the method of estimation.

Figure 3.1: Difference-in-Differences explained graphically


Outcome of interest
TB
Difference observed before the intervention
CB
Impact = difference between differences
TA
Difference observed after the intervention
CA
Before
After
= Treatment group
= Comparison group
= Outcome for treatment group if there had not been any treatment

Source: Authors

Exploiting discontinuity

Another possible way for estimation of the effect of a program is the *regression discontinuity design*. This method is applicable if a program has a certain threshold for eligibility (e.g. age, level of income, grade level, number of children, number of students in attending a school or other). If this is the case the program can be evaluated by comparing individuals who are *just above* and *just below* this administrative threshold. If the threshold is 15 years of age, the obvious control group is the group of kids who are 14 and hence do not receive the treatment.

It is important to note that our treatment group here is the group of those *just above the threshold*, in our example these are the 15 years old kids only. The problem of this approach is that it cannot tell us, unless you have some reasonable assumptions, what the effect of the program is on someone who is located comfortably above the administrative threshold. We say that we can estimate the *local average treatment effect* only; in our case it is the effect of the program on 15 years old kids.

Another problem with this approach is that the threshold must be strict. If administrators make exceptions on a non-random basis, this will bias results.

Other statistical methods

If none of the methods that we have described thus far are applicable, there are a couple more options for teasing out causal effects of a program. While, these methods are beyond the scope of this guidebook, you might consider consulting with someone who has an understanding of statistics in order to see if there are viable options.

Two common methods of identifying causality are:

- **Instrumental variables:** using a variable that affects assignment to the treatment group but *not* the outcome to estimate the effect of the treatment.
- **Propensity-score matching:** comparing observations with similar observable characteristics, differing only in assignment to the treatment, are compared.

Key reading

For more detailed information on other statistical methods, refer also to Duflo, Glennester, Kremer (2006): Using Randomization in Development Economics Research: A Toolkit. JPAL, MIT. Mimeo. Available online at:

http://www.povertyactionlab.org/sites/default/files/documents/Using%20Randomization%20in%20Development%20Economics.pdf

Note: the toolkit may be technically complicated in some parts, but most of the technical details can be skipped and the remaining information will still allow for designing of your evaluation.

Chapter 4: Comparing effectiveness and dissemination of results

- If there are more ways of achieving a similar goal in improving livelihoods, a cost-benefit analysis will help you to select the way that delivers most effect for least money.
- It is important the you do not forget to include the indirect costs and indirect benefits, otherwise you do not account for the entire effect and you might also have a hard time comparing different approaches.
- The results of an impact evaluation should be shared with other interested parties to increase the knowledge-base. This should be an obligation of every organization operating in development assistance.
- Proven impact of a particular project can significantly improve your chances for getting funding for a follow-up project using similar method.

In the previous chapters we have discussed why an impact evaluation may help you to understand whether your project works or not and—if properly designed—why it does or does not works. This might help you to improve your project design in the future or abandon an unsuccessful project if it fails to deliver results.

In this chapter, we will discuss some other ways in which this information might be beneficial for you, PIN or for the development community in general. Firstly, we will discuss how to compare cost-effectiveness of different projects and how to pick the one that delivers the most effect for the least money. Secondly, we consider how the results of an impact evaluation should be disseminated and how it might be beneficial in terms of obtaining further funding for your programs.

Selecting the most cost-effective program

When you carry out a proper impact evaluation your objective is to learn if the program has an effect a list of indicators that have been predefined. Aside for this, your objective should also be to find out which program would be the most cost effective in achieving your desired goal in case there are multiple ways of achieving it. In other words, you should want to know which program brings the most "bang for the buck." Figure 2.2 in Chapter 2 gives you a perfect example of why such thinking about program evaluation matters.

There are almost always more than one way to achieve to the same objective. This may be reducing pupils' absenteeism in schools, improving household diet, reducing incidence of HIV/AIDS or spreading knowledge about the benefits of animal vaccination. You should not only want to learn the effect of one particular program, but of all of them and do a comparison to pick the best one in the future.

It may seem that to do such evaluation of multiple approaches may be costly. Yes, this is true. Luckily, there are some funding agencies that give money for funding of your evaluation, such as the 3IE, the World Bank, Youth Employment Network by ILO, the UK Economic and Social Research Council (ESRC), the Evaluation Challenge and others. We report these sources in an Appendix to Chapter 2.

Also, there are lots of evaluation reports on the Internet, which may serve as your benchmark (although be careful that the source is trustworthy). Do you think your program can outperform the programs that have already proven to be effective? Then you may either compare your results to the existing results from previous studies or you may do evaluations of all the approaches yourself. The reason why you would do this is getting us back to the problem of *external validity*: different setting may deliver different results, unless you have convincing assumptions or theoretical reasoning why the results can be transferrable to other settings too.

If you want to conduct multiple evaluations during the running of one project, you may rely on the scale of the project. If you operate on an area that is large enough, you can randomize one part of the villages into one approach you want to evaluate, one part into the other approach, and the remaining part as a control group for both groups.

Through the previous chapters we mainly talked about the impact of the program. When assessing the cost-effectiveness or cost-benefit analysis, we also need to think of costs of designing and implementing the intervention. First, we will differentiate between the cost-effectiveness and the cost-benefit analyses and then we look at how to measure the costs related with the project.

- *Cost-effectiveness analysis (CEA)* relates costs of a program to specific measures of outputs or outcomes. In the examples given above, cost-effectiveness may be that we had to spend an additional $10 to get a child to stay in school for an additional month rather than staying home or working.
- *Cost-benefit analysis (CBA)* is a special case of CEA in which the indicators can be quantified in monetary terms, so that we can construct a real ratio of monetary benefits to the project participant to costs spent on him or her. CBA is usually considered when the programs you compare have multiple types of (potential) benefits and there is a consensus about how to quantify them in terms of money. (Source: Adapted from Hampel, Fiala, 2012)

Usually you would be able to do a CBA for programs such as a business training, training in income generating activities or in microfinance programs. Yet it is worth noting that, as we discuss in Bartos and Levely (2014) "Data collection", chapter 3, measuring individual income is especially difficult in the context of most developing countries where most of funds stem from very informal market transactions that are irregular and are almost never recorded.

In the following table we describe what is a cost and a benefit that should be accounted for and what should not be omitted in your own analysis.

Table 4.1: Costs and benefits of a program

Costs	Benefits
- **Direct resources.** Resources the program used directly for its purposes. If these are shared, you should know what share was used for this particular program. This includes staff salaries, stationery, travel costs etc.	- **Monetary benefits.** These are easy to quantify: income gains over the control group, increase of savings etc.
- **Capital spending.** This includes expenditures on cars, tools, computers.	- **Non-monetary benefits.** These cannot be expressed in terms of money. These might be changes in individual well-being, increased awareness about a particular problem such as HIV/AIDS transmission or similar.
- **Hidden costs.** This category is rather tricky, as it includes the so called opportunity costs to participants in terms of "what could have they done had they not visited the seminar", or time of volunteers expressed in terms of money that would otherwise be paid as salaries.	- **Spillover effects.** You should never forget about effects outside of the program participants. These might be positive or negative, but you should always take these into account.

Source: Authors and adapted from Hampel, Fiala (2012)

While the direct costs and direct monetary benefits are obvious candidates, the indirect effects are also crucial for a proper comparison:

- *Positive spillover effects*. All community members in a village where a farmer field school was established are likely to benefit from the knowledge gained by project participants. Every vaccinated individual or individual sleeping under a bednet reduces the likelihood of transmission of the disease for all people living in that area.

- *Negative spillover effects.* For example, training several tailors in an area where there are already a lot of tailors may drive prices so low that some of the tailors go out of business.

- *Displacement effects.* This effect occurs when, for example, a youth non-participant who would have found a job had the program not been implemented. A trader who would have sold the improved wheat seeds to a farmer had the free distribution of seeds by an NGO not taken place.

Even though these effects might be negligible in many cases, they should not be omitted in the discussion about the design of the impact evaluation and of a consequent CBA.

It should be clear that it is important that you keep track of all the categories presented in the table above separately and have as detailed data on it as possible so that you can do comparison across programs. Dhaliwal et al. (2012)[3]present a very detailed guide on cost-effectiveness methodology, where – among others – you can find more specific ways of quantification of costs and benefits.

Knowing the net benefits and net costs of the intervention, it is then possible to calculate the ratio of benefits and costs, which can be compared across programs. For example, the benefits/cost ratio is 2:1 if net benefits of the program are $200 per person and net costs are $100. Once again, it is important to remind here that you should be comparing the same set of costs and the same set of benefits across all programs you compare, i.e. to take into account all categories we discussed above. Then, you should want to put most effort to proposing the program that is most cost-effective.

Dissemination of results

The results of your impact analysis should not result in a single conclusion of 'success' or a 'failure' on an evaluation report. The question you have asked might raise new questions for a continuous journey towards understanding of what works best in development assistance.

Share the results of your impact evaluation. It is also important to share the information with the outside world. There are two reasons for doing this. The first one is that the NGOs should work jointly on an effort of building a knowledge-base that helps to improve the assistance. Usually the outputs of an impact evaluation are:

- **Evaluation report.** Detailed report introducing background of the project, question studied, method used, discusses both external and internal validity of the results and provides policy recommendation.
- **Policy brief.** You should prepare a short report summarizing the project and its results for a broad audience.
- **Country presentations and workshops.** Presenting the results to the local audience has most potential for further use of the results. The flow of information may be in both directions, as the local stakeholders might help you interpret some unexpected results. You can also expect that the local audience might benefit most, as the results are easier to replicate in a similar cultural environment.
- **International conferences.** The ultimate goal is to share the results with the global development community. This is also a good testing-ground for comparison of cost-effectiveness of the approach you propose. *(Source: Adapted from World Bank, 2012)*

Promote your results to obtain donor funding. The information, however, is also a valuable asset for your organization. The rigorously tested results might significantly increase your chances of getting funding for your future projects.

To inform the donor successfully using the findings from your evaluation, you should:

1) Identify the goals of the call you want to submit your proposal to

[3] Dhaliwal, Duflo, Glennerster, Tulloch (2012). Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries: A General Framework with Applications for Education. Abdul Latif Jameel Poverty Action Lab, MIT. Mimeo.

2) Select the relevant data from your evaluation that shows that you are successful in addressing this mission; and
3) Tailor your presentation to the funding source's interests and purpose.

Imagine you are a funding agency comparing one proposal, which “expects to deliver achievements” and one, which “has proven impact of the program and wants to deliver the same benefits to broader population”. Obviously, the latter is the candidate for funding.

Failed projects should also be reported. Keep in mind that a report of a failed program is also important, especially if the evaluation provides the answer to the most important question of why it failed or for which part of the population it was unsuccessful. This information is important for other organizations, which might otherwise start a similar project only to fail again.[4]

Key reading

For more detailed information on result’s dissemination and impact evaluation in general, you might want to consult Module 7 of the World Bank’s Impact Evaluation Toolkit: Vermeersch, Rothenbühler, Sturdy (2012): Impact Evaluation Toolkit: Measuring the Impact of Results-Based Financing on Maternal and Child Health. World Bank.

For detailed guide on cost-benefit analysis, consult JPAL’s guide: Dhaliwal, Duflo, Glennerster, Tulloch (2012). Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries: A General Framework with Applications for Education. Abdul Latif Jameel Poverty Action Lab, MIT. Mimeo.

[4] In academia, publications that fail to find a significant result often remain unpublished. This leads to a so-called *publication bias*.

Impact evaluation: A practical guide to designing and administering impact evaluations of PIN programs

Vojtěch Bartoš and Ian Levely