CHAPTER 2

The Spectrum of Impact Evaluations

Abstract This chapter underscores the importance of causal attribution and takes the readers through various impact evaluation methodologies that enable evaluators to measure the causal impact of policies. Using case studies, it highlights important assumptions, advantages, and disadvantages of each methodology to give readers a sense of how these techniques can be applied to issues of sustainable development.

Keywords Attribution • Causal analysis • Counterfactual • Case studies

Policy-makers are increasingly seeking answers to the question of what works and what does not in addressing issues of making development more sustainable. Crucial to answering these questions is the ability to show, to the extent possible, attribution, that a change in the outcome, say a decrease in air or water pollution, is causally linked to a policy under consideration.

Conventional thinking is that the scale of issues such as climate change and income inequality is too large for them to lend themselves naturally to impact evaluations (IEs). Cameron, Mishra, and Brown (2016) did a systematic review of published IEs of international development interventions. Out of 2259 studies they reviewed (between 1981 and 2012, though the number of studies increased significantly after 2008), 1476

V. Thomas, N. Chindarkar, *Economic Evaluation of Sustainable Development*, https://doi.org/10.1007/978-981-13-6389-4_2

were on health, nutrition, and population; 521 were on education; and 341 were on social protection. Only 14 were on energy and 124 on environment and disaster management. The systematic review also finds that over the years there is stagnation on rigorous IEs on economic policy, energy, transportation, and urban development.

Given the complex nature of issues of sustainability, skepticism about their suitability for evaluation is justified. How can we randomly assign deforestation or air pollution to treatment and control areas? How can policy-makers experimentally roll out policies that are aimed at bridging rural-urban gaps? Is it politically feasible to provide social mobility opportunities to some but not to others? In this chapter, we discuss using real-life policy case studies how econometric experimental and quasi-experimental IE methods can be extended to overcome practical and empirical challenges.

Why Impact Evaluation?

Public policy-makers are guided by goals, objectives, and indicators. IE not only assesses whether goals were reached but also helps to understand the mechanism by which the impacts were generated. The shift toward evidence-based policy-making calls for a good understanding of what IE can and cannot do as well as how it can be designed, applied, and replicated. The core objective of IE is to assess how much of an impact can be attributed or causally linked to a specific project, program, policy, or even a shock such as a climate-related natural disaster.

The application of IE is not limited to small-scale and targeted projects. IE tools offer flexibility to evaluate targeted projects such as the impact of change in classroom pedagogy in public schools on children's learning outcomes, as well as to evaluate large-scale, national-level programs and policies such as compulsory education or subsidies for the education of girls. It thus has the capacity to assess the impact of interventions related to an array of issues included in the SDGs such as poverty alleviation, social inclusion, and environmental stewardship.

To illustrate the practical and empirical challenges in conducting IEs of sustainable development policies, let us consider, as an example, the policy to control illegal deforestation in Brazil, which is directly related to SDG 12 (climate action). We know that deforestation activities tend to be concentrated in areas with high levels of forest resources and low levels of governance and monitoring. Therefore, policies to control deforestation

are often geographically targeted. One such policy is the Priority Municipality (PM) program introduced in the Brazilian Amazon in 2007 that instated rigorous monitoring in areas that experienced extensive illegal deforestation (Slough and Urpelainen 2018). While this is the typical policy response, assessing the success of such a policy in curbing deforestation can be challenging.

An obvious challenge for evaluators is that the PM program does not have control over the movement of extractors engaged in illicit deforestation from the priority areas to other areas where there is no restriction. It is thus possible that extractors who previously practiced illegal logging in priority areas decided to move elsewhere and continue their activities after the program was introduced. If an evaluation only accounts for the changes in deforestation within the priority areas and fails to consider the negative spillovers in other areas, the results may suggest a significant decrease in deforestation rates. This might lead to the conclusion that the program achieved its objectives although there may be serious negative spillovers in other areas.

The question IE should ask therefore is "What is the *treatment effect* of a program on an outcome?" Answering this question requires a good understanding of causal inference and the spectrum of available evaluation methods so that the most suitable one can be chosen.

Causal Inference and Counterfactuals

In seeking answers to questions about the effect of an intervention, the challenge is to establish causality between a program and an outcome. Econometric IE is a tool that helps us empirically establish causality by measuring the differences (Δ) in outcome (Y) of the program participants (T) and outcome of the nonparticipants (C), given by the formula:

$$\Delta = (Y|T) - (Y|C)$$

To further illustrate the complexity in establishing causality, let us look at another example. Investing in rural infrastructure such as electrification and roads is considered vital to reducing rural-urban inequality and promoting inclusive growth (UN 2016). Chen, Chindarkar, and Xiao (2019) examine the causal effect of an electrification upgrading program on improvements in rural health systems including health services utilization,

health information, and health facilities. In 2003 the state government of Gujarat, India, launched the Jyotigram Yojana (JGY), which provides 24-hour, high-quality electricity to rural areas. Stable electricity supply is an enabler of universal access to health care as it mediates health services utilization such as child immunization and ante-natal care, access to health information through electronic media, and functioning of health equipment in rural health centers (Chen et al. 2019; WHO 2014).

Therefore, improving the quality of electricity supply can be expected to improve rural health systems. This would effectively result in greater health equity as the health gap between rural and urban households is narrowed. Considering just one of the outcome indicators—child immunization—the formula indicates that the gap between the immunization rate of children from households that reside in villages that were electrified under the program $(Y|T)$ and the immunization rate of children from households that reside in villages which remained unelectrified $(Y|C)$ is the effect caused by program (Δ). An important question is whether the households in electrified and unelectrified villages are comparable.

In order to draw a causal inference, the observed outcome of the treatment group—or the individuals or households affected by the program—needs to be compared with the potential outcome of the group had they not been exposed to the program. This is referred to as the counterfactual outcome. The difference between the actual outcome and counterfactual outcome can be attributed to the program because in this scenario the two groups are identical in expectation except for their treatment status.[1] In other words, we expect the two groups to be identical, on average, in the absence of the program. In reality, however, the counterfactual outcome is not observed. It is not possible to observe the immunization rate of the same children with and without electrification simultaneously. The counterfactual thus needs to be estimated, and this is where econometric tools come in handy.

Estimating the Counterfactual

The identification of program impact requires generating two groups that are statistically identical in expectation in the absence of the program: one group affected by the program, called the treatment group, and a group not affected by the program, called the control group. Comparing outcomes of these groups ensures that the difference between the outcomes of

the treatment group and the control group is due to the program. The challenge in identifying the causal impact is to find a valid control group.

In the case of rural electrification, the first thing that one would think of is to compare the immunization rate before the treated households were exposed to the program and after the program was implemented:

$$\Delta = (Y_{t1} | T) - (Y_{t0} | T)$$

As illustrated in Fig. 2.1, the immunization rate before the intervention, which is an estimate of the counterfactual, is Y_{t0}, and the rate after the intervention is Y_{t1}. The before-and-after comparison seems to suggest that the intervention increased the immunization rate by $Y_{t1} - Y_{t0}$.

However, consider the case where the majority of rural households suffered from a drought. A decline in income from the crop damage could have discouraged parents from investing in their children's health. Then the outcome in year 1 in the absence of intervention would likely be lower than Y_{t0}. In this case, the actual impact of the program might be $Y_{t1} - Y_{t1'}$, which is larger than $Y_{t1} - Y_{t0}$. A failure to account for the effect of drought will result in underestimating the impact.

On the other hand, other factors might positively affect the health outcomes of children over time such as an increase in household income or

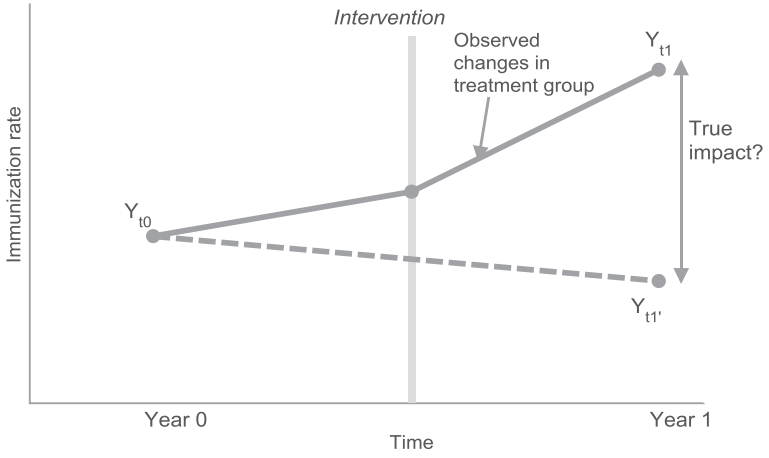

Fig. 2.1 Before-and-after comparison. (Source: Authors' illustration)

increase in health budgets allocated to the state. Ignoring these factors that might lead to an increase in the child immunization rate could result in overestimation of the impact of the program. The same considerations hold for other types of interventions, such as education, vocational training, and micro credit. The baseline outcome can hardly serve as an accurate measure of the counterfactual.

With such situations in mind, another method of counterfactual estimation uses a group that is not exposed to a program. As shown in Fig. 2.2, the gap in observed outcomes of the treatment group and the control group is $Y_{t1} - Y_{c1}$. This estimates the impact of the program only if we can assume that the *changes* in the immunization rates caused by the electrification program would not be different for the two groups. In many cases, the program is targeted toward areas or groups of people who are in need of the program. If the households in target areas of rural electrification program are different from households in non-target areas, say in terms of their socioeconomic conditions, then we are essentially comparing apples to oranges.

The effect of an intervention is then likely to be different for the treatment group and the control group. If the counterfactual of the treatment group were observed, then we would know that the real impact is observed

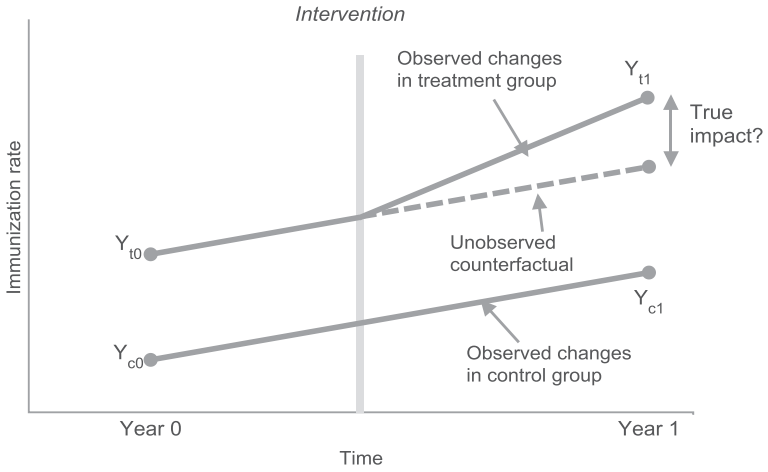

Fig. 2.2 With-and-without comparison. (Source: Authors' illustration)

changes in treatment group ($Y_{t1} - Y_{t0}$) less observed changes in control group ($Y_{c1} - Y_{c0}$). In other words, the change in outcome of the control group is the change that would have occurred without any policy intervention and therefore that part cannot be attributed to the intervention.

Another important factor that could taint the estimate of counterfactual is the selection mechanism. Information on the electrification program could attract some households to move from villages with poor electricity supply to the target villages. Those households might also have higher health awareness than those who stay in non-targeted villages. Naturally, the effect of the program is larger for such households compared to others, which again leads to a bias in the program's impact estimation.

Selection bias occurs when the program participation decision is correlated with unobserved factors. This is a serious concern in various types of interventions. In a conditional cash transfer program where the cash is provided on the condition of children being in school, it is highly likely that parents with higher motivation to send children to school, which is typically unobserved, will participate in the program. Depending on the selection mechanism, a simple with-and-without comparison could bias the estimated impact of the program.

Establishing a Theory of Change

Before applying either experimental or quasi-experimental methods, evaluators must lay out the theory of change, which is the causal logic of how and why a particular program is likely to achieve the intended outcomes. It is guided by existing theoretical and empirical literature and helps to build appropriate hypotheses for the IE. For instance, the theory of change underlying improved rural electrification and child immunization may be that electrification provides better vaccine storage facilities. There might also be positive spillovers from improvements in other aspects of the health system such as improved health facilities and increased access to health information.

In the IE, the econometric analysis would examine the effects of improved electricity access on child immunization rates as well as mediating factors such as health facilities and health information.

Internal Validity and External Validity

IE seeks answers to the causal effect question, and its usefulness greatly depends on its validity, both internal and external. For an evaluation to have internal validity, the outcome of the control group needs to be indeed a valid counterfactual and the estimated impact needs to be solely attributed to the program. In some cases, explanatory variables may not be well specified (referred to as omitted variables) or accurately measured (measurement error). There might be issues related to program assignment (imperfect compliance with the treatment or correlation between treatment assignment and outcome). Other factors such as attrition and externalities also put internal validity at risk. Any of these might cause low internal validity, undermining the inference of causality.

External validity means that results are applicable or generalizable to different populations, contexts, and outcomes. The threats to external validity essentially concern important interactions between the treatment and individual characteristics, location, or time (Meyer 1995). The less the likelihood of violating external validity, the more confident policy-makers can be in applying the impact evaluation learning to populations beyond the one under examination, or to other contexts.

While strong validity improves the quality of IE, incorporating IE in policy design a priori could help minimize threats to internal and external validity. Prospective IE can be incorporated in the process of policy design so that valid counterfactuals and data are available in the future.

Random Assignment

It is not possible to avoid all threats to internal and external validity, but there is a tool to help deal with it: random assignment. Often referred to as randomized controlled trials (RCTs), random assignment is increasingly used in economics and other social sciences. RCTs give every eligible unit an equal probability of being selected into a program. Such a selection mechanism not only generates a valid counterfactual, but it is also transparent and accountable. RCTs are often viewed as the most credible approach to establishing causality because they require few statistical assumptions and analysis can be done using simple econometric methods.

Random assignment of treatment and control groups produces two comparable groups when sample size is large. This is based on the property called the law of large numbers (LLN). The LLN states that a sample

average will approximate the average of the population from which it is drawn as the sample size grows larger. The gap between averages of two groups can then be interpreted as the unbiased estimator of the average treatment effect (ATE). This can be expressed as

$$Y_i = \alpha + \beta T_i + \varepsilon_i$$

where T_i is the treatment status dummy that equals 1 if a randomly selected unit, i, is treated, and 0 otherwise. Random assignment ensures that T and i are independent and the estimated treatment effect $\hat{\beta}_{\text{OLS}}$ is unbiased.

Sampling and Validity Issues in Randomization

In practice, it is not simple to generate two groups that are the same except for the treatment status. Random assignment is commonly conducted in two steps. The first step is to randomly select a sample of potential participants from the eligible population. The second step is to randomly select units to be assigned to treatment and control groups. Each step ensures external and internal validity, as in Fig. 2.3.

Although RCTs seem to be a solution for establishing validity, some have pointed out that in reality they might be compromised (Deaton and Cartwright 2016; Ravallion 2018). Internal validity of estimates could be

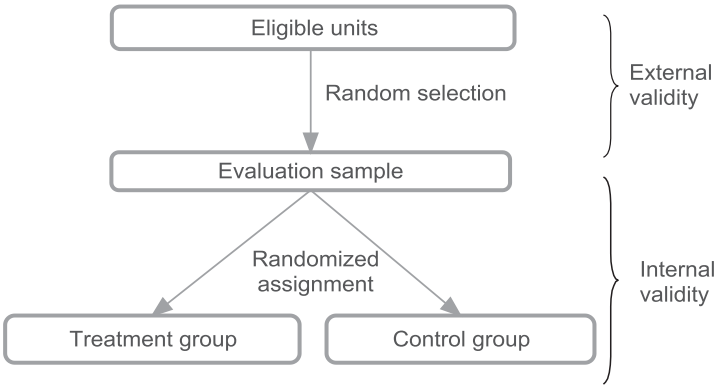

Fig. 2.3 Random sampling and randomized assignment of treatment. (Source: Authors' illustration)

put at risk when compliance with treatment assignment is not perfect, there are externalities, or randomization is conditional on observed variables. Most programs aim to reach all members of the randomly assigned treatment group. In many cases, however, full compliance with the treatment assignment may not be achieved. This could be due to the behavior of both the treatment and control groups.

Take, for example, a vocational training program offered to randomly selected schools. Some in the treatment schools who are offered a free training course may not be motivated to take up the program. On the other hand, some in the control group may decide to transfer to a school in the treatment group. These behaviors change the original treatment assignment status and contaminate the randomized design.

A second threat to internal validity is externalities. Most social science RCTs are conducted in the field, where externalities are often generated, and not in a laboratory. In the vocational training program example, consider a case where there are two friends, one assigned to the treatment group and the other assigned to the control group. It is conceivable that the one in the treatment group performs well owing to the training program and passes on information to the friend in the control group, who, because of the information received, also performs well.

A third threat is imperfect randomization. Randomization is often conditional on a set of observed variables. The assumption that, conditional on the observed variables, the potential outcomes of treatment and control groups are identical in expectation could eliminate the selection bias. For this assumption to hold, the set of observed variables needs to include all the relevant variables that account for the differences between the two groups. Incomplete data on observables could result in selection bias owing to omitted variables. It is, therefore, important to carefully select the variables according to the setting and purpose of the program.

While randomization can eliminate selection bias to a large extent, it does not guarantee that findings from an RCT in one context will necessarily hold in others. Duflo, Glennerster, and Kremer (2007) discuss that there are three major factors to consider in examining the external validity of RCT results. First is careful design and documentation of the intervention. While an RCT can be well designed and successfully implemented as a pilot or on a small scale, scaling up and replication can present a bigger challenge, particularly when the evaluation design is not clearly documented. Procedures must be put in place to record the study design and implementation processes so that policy-makers can use them for program expansion and replication.

A second factor is whether findings from an RCT on one sample can be generalized to the population. As previously discussed, external validity can be strengthened by randomly selecting the sample from the eligible population in the first step and randomly assigning the treatment and control groups in the second step. However, in practice, sampling may not be random and therefore not representative of the population. In some cases, the eligible sample may be selected because it is convenient. The sampling decision is often guided by the availability and approval of study partners such as nongovernment organizations or local governments. This severely constrains the generalizability of RCT findings to samples beyond the one being studied.

A third factor is how the effect of the program would differ if the treatment were slightly different. In a conditional cash transfer program, what would happen if the amount offered were increased? Would the results change if the age of eligibility were lowered? Conducting RCTs with multiple treatment arms could offer insights into what works and what does not and how the program could be tweaked. However, this increases the design and implementation complexity. A sound approach is to have an appropriate theoretical framework to judge which treatment arms are important.

Finally, RCTs may not always yield estimators that are more unbiased relative to observational or nonrandom studies. The first reason for this is linked to issues of external validity and choice of samples for RCTs (Ravallion 2018). The second reason pertains to the variance of errors in estimates from RCTs compared to observational studies. Ravallion (2018) argues that, despite the bias, the variance of errors from observational studies that use large sample sizes could be low enough to assure that they are closer to the true population parameter. In contrast, despite the lack of bias, the small and selective samples that are often used in practice for conducting RCTs may yield estimates that are further from the true population parameter.

Which Treatment Parameters Are of Interest?

Deaton and Cartwright (2016) suggest there are three alternatives when internal validity is violated and an ideal ATE, β, cannot be calculated. First is to calculate the difference in means between those who, regardless of their assignment status, received the treatment, β_1, and those who did not. This is called the average treatment effect on the treated (ATT).

Second is to estimate what is called "intent to treat" (ITT), β_2, which is the difference between the average outcome of those who were intended to be in the treatment group and those who were intended to be in the control group, according to the original treatment assignment and regardless of whether they complied. The ITT estimate will be different from β unless there is perfect compliance. Perfect compliance is often violated in field experiments, and those who do not comply with their assignment status tend to have different characteristics compared to those who comply, making β_2 a parameter of interest.

Third is an estimator, β_3, called the local average treatment effect (LATE). In many cases, the program is not directly offered to all individuals. Rather only information on the program is randomized, and individuals select themselves into the program based on the information. Such experimental design is common in social programs offering vocational training. LATE estimates the program effects for the subgroup that complies (p), and it is calculated as $\beta_3 = \beta_2 / p$. In particular, it only accounts for those whose treatment status was induced by the randomized program information. In other words, it is the average causal effect for those who participated in the vocational training program only because they were offered information without which they would not have participated.

These three estimators are average over different populations; therefore, they are different without additional assumptions on the heterogeneity of treatment effects (Deaton and Cartwright 2016). In general, it is natural to assume there are different characteristics for those who comply with the treatment assignment and those who do not. For instance, those who are offered information but decide not to participate in the vocational training program may already have high skills and not feel the need for further training. Those who participate even if they are not offered information may have higher motivation and may learn more from the same training than participants in the treatment group. Given that treatment effects are often heterogeneous, it is vital to be clear about what is being evaluated and which treatment parameter is being estimated.

Need for Baseline Information

Since the treatment assignment may not be completely random even in RCTs, a good practice is to conduct baseline surveys to examine initial conditions as well as their interactions with the impact of the program. Baseline surveys are crucial in conducting balance checks. Balance checks

enable evaluators to statistically judge whether the treatment and control groups are similar on average before the intervention is introduced. These can be performed using simple hypothesis tests of difference in two sample means. The expectation is that there should be no systematic differences on average in the observed characteristics of the treatment and control groups. This strengthens both the internal and external validity of the findings from the RCT.

Additional balance checks can be performed in case of attrition, where units assigned to the treatment or control group drop out of the experimental study. Here we would compare the balance between the treatment and control groups before and after attrition. Again, the expectation is that there are no differences between the treatment and control groups post attrition, meaning attrition was not systematic and therefore should not be a validity concern.

Quasi-experimental Methods

Can policy-makers randomly select where to construct a new road, build irrigation systems, or supply electricity? RCTs have many advantages; however, public policy implementation rarely follows experimental design, as it may not fit with program objectives, be costly, or be politically unfeasible or unethical. In circumstances where randomization is not feasible, it is possible for policy analysts to exploit natural experiments or quasi-experiments that offer opportunities to select a control group that was excluded from the program but shares similar characteristics with the treated group.

Various econometric tools are available to identify causal effects using quasi-experimental methods. Each method comes with a different set of assumptions and data requirements that need to be considered carefully. Each has its advantages as well as limitations. In this section, we will discuss four commonly used quasi-experimental methods—difference-in-differences, regression discontinuity, instrumental variables, and propensity score matching.

Difference-in-Differences

Interventions to tackle environmental issues often adopt geographical targeting as policy needs to prioritize areas with more severe environmental deterioration. The study of deforestation policy in the Brazilian Amazon is

a case when randomized policy implementation is not possible because, by its very nature, priority areas are located where deforestation activities are more extensive (Slough and Urpelainen 2018). This program which introduced rigorous monitoring in areas with extensive deforestation could generate a displacement of deforestation to neighboring areas if there is limited state capacity to properly implement the program.

Evaluations if designed properly can help examine the effect of policies beyond the targeted areas. To create a natural experiment setting, that is, assignment of priority areas, the study uses changes in priority areas over time according to changes in deforestation rates. The study combines information on priority areas designated by the government and from satellite monitoring data to identify forest clearing. It then exploits the variation in designation of priority areas over time to evaluate the impact on deforestation in the target areas as well as neighboring areas using the difference-in-differences (DID) method.

A simple before-and-after or with-and-without comparison would not give an accurate causal estimate of such a program. The quasi-experimental setting always leaves concern about nonrandom program implementation. If both pre-program and post-program data are available for both treatment and control groups, a method called DID is one way to eliminate the bias. DID makes use of these data to obtain a valid counterfactual to estimate the effect of an intervention or a program by comparing the average change in outcome over time between the treatment group and control group.

Using the example of deforestation policy, average change in the probability of a deforestation event (the outcome variable of interest) in the priority area and control area is illustrated in Fig. 2.4. Before the program, the average probability of a deforestation event for the treatment group is *A*, which is higher than the average probability for the control group, *C*. The average change from year 0 to year 1 in the control area is the counterfactual for the priority area. This means that in the absence of the program, the priority area would follow the same trend as the control area and reach *E* in year 1. This decrease from *A* to *E* needs to be subtracted from the actual change in the treatment group, from *A* to *B*.

Therefore, as summarized in Table 2.1, the impact of the program is calculated as

$$\text{DID} = (B - A) - (D - C) = (0.76 - 0.86) - (0.68 - 0.70) = -0.08$$

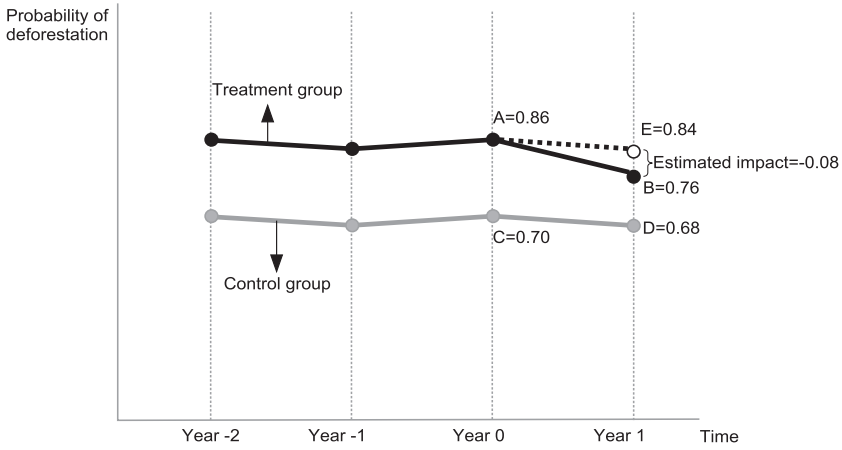

Fig. 2.4 DID applied to deforestation policy. (Source: Authors' illustration)

Table 2.1 Calculating the impact in DID method

	After	*Before*	*Difference*
Treatment group	B	A	B – A
	0.76	0.86	−0.10
Control group	D	C	D – C
	0.68	0.70	−0.02
Difference	B – D	A – C	(B – A) – (D – C)
	0.08	0.16	−0.08

Source: Authors' illustration

As DID averages the treatment effect over the entire treatment and control population, the resulting estimate is the ATE. Econometrically, estimation of the ATE using DID is done using the following regression model:

$$Y_{it} = \beta_0 + \beta_1 TREAT_i + \beta_2 POST_t + \beta_3 TREAT_i * POST_t + \varepsilon_{it}$$

where Y_{it} is the outcome variable of interest. $TREAT_i$ equals 1 if treated and equals 0 if not treated. $POST_t$ equals 1 if post-program period and equals 0 if preprogram period. β_3 gives the DID estimate.

The DID Parallel-Trends Assumption

DID provides an unbiased estimate of the treatment effect under the parallel-trends (or equal-trends) assumption. The parallel-trends assumption is that in the absence of the program, the difference in the outcome over time for the treatment and control groups would follow the same trend or the outcomes would move in tandem. It is important to note that the assumption does not state that in the absence of the program the outcome for the treatment and control groups would be the same. Rather, it assumes that the outcomes, although different, follow the same trend.

Although there is no formal statistical test to prove that both groups follow the same trends in the absence of a program, there are several ways to check the validity of this assumption. First is to graphically compare the trends in outcome using several periods of preprogram data. If the prepolicy trends are the same for the treatment and control groups, then it is safe to assume that they follow parallel trends. Figure 2.4 shows that control and treatment groups follow the same trends before the program implementation. This gives more confidence in assuming that the two groups would follow the same trends after year 0 if not for the program.

A second way to test the validity of the parallel-trends assumption is to perform what is known as a placebo or a falsification test using a different slice of time, a different sample, or a different outcome variable. The idea is that the program should have no impact on this differently chosen time, sample, or variable. If we do find significant effects, then there might be some unaccounted-for or unobserved factors outside of the program that caused the changes in outcome. Slough and Urpelainen (2018) use the 12-month time period prior to actual priority area assignment and run their DID model only on pre-program data. The hypothesis is that there should be no significant reduction in the probability of deforestation during this period. Their placebo test results support this hypothesis, and therefore the parallel-trends assumption holds.

Advantages and Limitations of DID

In many quasi-experimental programs, the treatment assignment rules are not as clear as in experiments. The advantage of DID is that it controls for unobserved as well as observed characteristics that affect participation in the program as long as the characteristics are time invariant. Many observed characteristics, such as geographic and climatic conditions, or unobserved characteristics, such as a culture of conservation, are likely to be constant over time. DID's biggest limitation, however, is that it does not control

for time-varying unobserved characteristics, like the ability and motivation of local government personnel in implementing deforestation policy. If different personnel are in charge at different points in time, their ability and motivation to implement the policy with stringency is likely to be time varying. Therefore, we might still estimate slightly biased treatment effects.

Regression Discontinuity Design (RDD)

Often public policies follow eligibility criteria for targeting purposes. Common examples of these are pension programs, which impose an age eligibility criterion, or poverty-alleviation programs, which impose a minimum-income criterion. These criteria can be exploited to create comparable treatment and control groups and to evaluate large-scale programs. The example of an evaluation by Chen et al. (2013) of energy policy in China can help illustrate this. The study applies a quasi-experimental method called regression discontinuity design (RDD) to evaluate an energy program that provides coal for winter heating in Northern China.

RDD can be used to evaluate programs that have a continuous eligibility index with a clearly defined eligibility threshold or cutoff. The observations close to the cutoff are divided into the eligible (treatment) and non-eligible (control) groups, and their outcomes are compared in order to estimate the local average treatment effect. As RDD restricts the treatment and control groups only to a certain bandwidth around the cutoff to ensure that they are similar on average, the treatment effect cannot be generalized to the entire population. We are therefore only able to estimate the LATE.

During the period of central planning (1950–1980), the Chinese government provided free coal for winter heating to homes and businesses as a basic right in Northern China. Such coal combustion releases harmful air pollutants that are known to adversely affect human health. Owing to budgetary limitations, the free provision was restricted only to areas north of the Huai River (shown in Fig. 2.5). This created a quasi-experimental opportunity to compare the cardiorespiratory mortality rates and life expectancy of the treatment group, residing just north of the river that received free coal, and the control group, residing just south of the river that did not receive free coal. Here the distance from the river is the continuous eligibility index, and the river itself is the spatial cutoff point. As the two groups reside within close proximity to the river, they are assumed

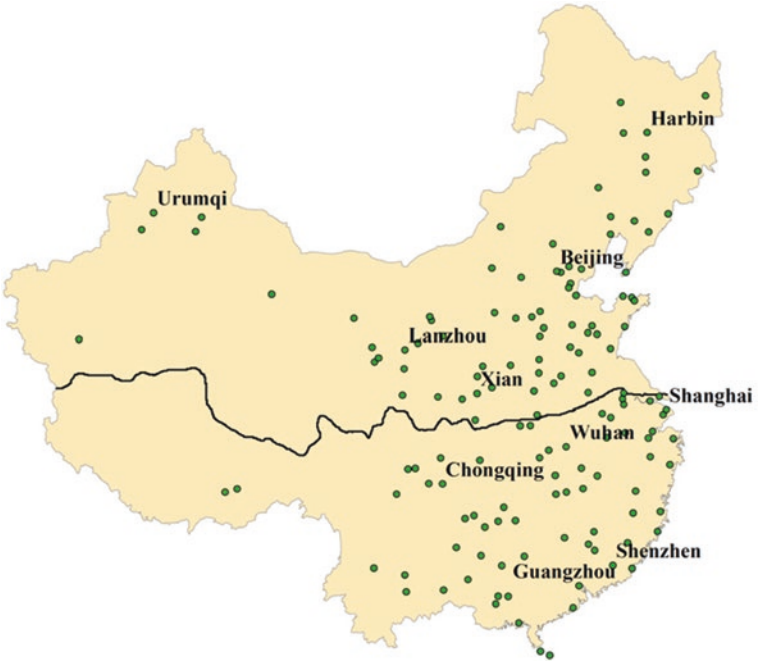

Fig. 2.5 Cities to the north and south of the Huai River. (Source: Chen et al. 2013)

to be similar in all important aspects except for the amount of pollutants they were exposed to.

The RDD treatment effect can be estimated using the following linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + \rho w_i + \varepsilon_i$$

where Y_i is the value of the outcome for unit i, in this case, life expectancy at birth; x_i is the continuous eligibility index, in this case, the degrees north of the Huai River; w_i is the dummy variable that indicates whether the unit is in the treatment or the control group, in this case, 1 for locations north of the Huai River and 0 otherwise.

The study finds a striking decline in life expectancy north of the Huai River. Figure 2.6 indicates that average life expectancy at birth reduced by

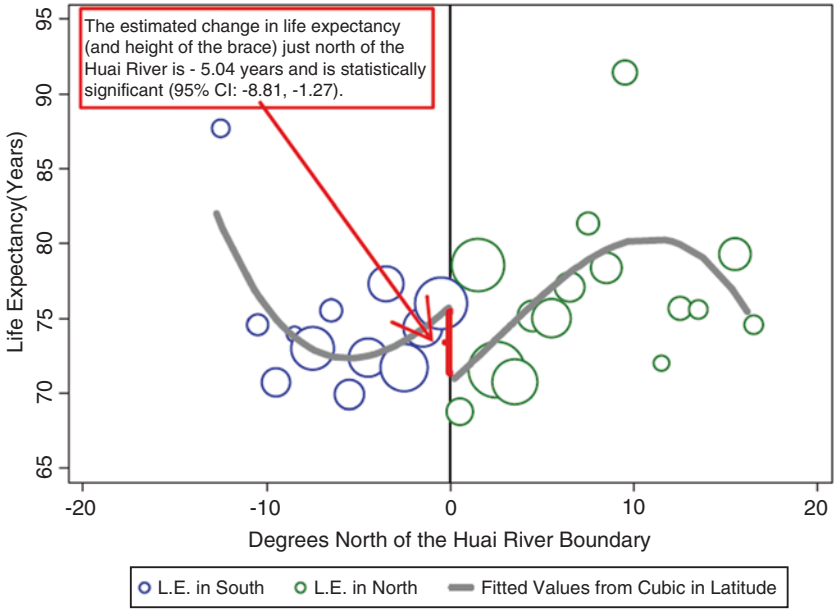

Fig. 2.6 Fitted values from RDD estimation. (Source: Chen et al. 2013)

almost five years for those living just north of the river owing to increased exposure to air pollution.

This study demonstrates the difficult trade-off between economic growth, public health, and environmental quality that many growing economies face today. The level of pollutants at the time of the study could be used as a reference for cities in developing countries such as Brazil and India where pollution is a serious issue. Though certain adjustments are required in applying the findings in different countries or different contexts, the insights obtained in such evaluations are useful in controlling or avoiding interventions that might have negative consequences on the environment.

Advantages and Limitations of the Regression Discontinuity Design (RDD)

The RDD method exploits the opportunities naturally generated by the program eligibility criteria and allows unbiased estimates of the treatment

effect. An advantage of the RDD method is that it does not require any eligible units to be untreated for the purposes of the IE. The treatment effect, however, is valid only for the units around the eligibility cutoff. In other words, the estimated treatment effect is the LATE. Therefore, an important limitation of RDD is that the estimated effects may not always be generalized to units whose eligibility scores are far from the cutoff point.

A further challenge arises when the enforcement of eligibility is not clear-cut or "sharp" but is "fuzzy." This means that not all eligible units may be affected by the program, and some ineligible units might be affected. If the compliance with the eligibility criteria is "fuzzy," the eligibility score can be replaced with a probability of participating, and the estimated treatment effect is the difference around a neighborhood of the cutoff score.[2]

The statistical power of analysis presents another challenge that arises because the RDD method only estimates impact around the cutoff. This restricts the number of observations used to estimate the impact, which lowers the statistical power of analysis. The bandwidth around the cutoff point needs to be determined so as to include a sufficient number of observations while maintaining the balance of important characteristics to make the treatment and control groups comparable.

Finally, problems also arise when it is possible for participants to manipulate eligibility criteria. For instance, if corruption is high, it may be possible for people to provide fake documents to make them eligible for the program. This contaminates the quasi-experimental features of RDD and produces biased estimates. A simple way to identify manipulation is to plot a histogram of eligible units along the continuous eligibility criteria. The appearance of far too many units clustered just above the eligibility criteria might indicate potential manipulation, and policy analysts will need to dig further into how the program was implemented on the ground.

Instrumental Variables (IV)

Another important quasi-experimental method is called instrumental variables (IV). As discussed previously, there might be a systematic correlation between program participation and unobserved characteristics of the participants, in what is often referred to as endogeneity. Endogeneity may arise if participants self-select themselves into a program or they do not comply with randomized experimental design or program eligibility criteria. IVs allow us to address such issues of endogeneity.

Let us understand IV using IE of the Moving to Opportunity (MTO) program, an experimental housing-mobility program introduced in 1994 in several cities in the United States. This program was motivated by the fact that there is significant geographical disparity in social and economic status and was implemented to examine whether moving from a high-poverty neighborhood to a low-poverty neighborhood improves social and economic prospects of low-income families. Under MTO, eligible families were randomly assigned housing vouchers by the US Department of Housing and Urban Development to move from poorer neighborhoods to better-off neighborhoods. They were also provided counseling services to adjust to the new neighborhoods. The control group families did not receive any vouchers.

However, they continued to receive other government assistance they were eligible for. As discussed in the randomization section, in reality, perfect compliance with the randomized treatment assignment is rare. In the case of MTO as well, not all families who were offered the housing vouchers actually took them up. The evaluation of MTO conducted by Chetty, Hendren, and Katz (2016) applies IV to address this imperfect compliance in voucher take-up.

Without full compliance, the estimated treatment effect is either that of offering a program (ITT), that of participating in the program (ATT), or limited to those who complied with the experimental design or program eligibility criteria (LATE). As previously discussed, the basic ATE estimation regression setup is expressed as follows:

$$Y_i = \alpha + \beta T_i + \varepsilon_i$$

When treatment assignment is not random in reality, treatment dummy T and the error term ε are systematically correlated, that is, $\text{cov}(T, \varepsilon) \neq 0$. The IV method aims to remove this correlation by isolating the variation in T that is uncorrelated with ε. For an instrumental variable, Z, to be valid, it must satisfy the following two conditions:

$$\text{cov}(Z, T) \neq 0 \text{ and } \text{cov}(Z, \varepsilon) = 0$$

The first condition is called relevance, and it shows that an IV is correlated with the treatment variable. The second, called exogeneity, shows that the IV is uncorrelated with the error term. Essentially, we rule out any

direct effect of the IV on the outcome or any effect coming from unobserved or omitted variables. This is also known as exclusion restriction.

Chetty, Hendren, and Katz (2016) use the randomly assigned MTO treatment indicator as an IV (Z) for actual take-up of housing vouchers. As the random treatment indicator is correlated with treatment assignment and uncorrelated with the error term, it satisfies the IV validity conditions. The exclusion restriction is that the MTO voucher offers affect the outcomes only through the actual use of the voucher. They use a two-stage least-squares (2SLS) regression that is composed of two regressions.[3] The first stage regresses the voucher dummy variables on the random treatment indicator Z, additional covariates, and the error term, u_{1i}:

$$T_i = \pi_0 + \pi_1 Z_i + u_{1i}$$

Because Z_i is uncorrelated with u_{1i}, the estimate of π_0 and π_1 is uncorrelated with u_{1i}.

The second stage regresses the outcome variable on the predicted value of voucher take-up from the first stage with other covariates and the error term:

$$Y_i = \alpha_0 + \lambda \widehat{T}_i + u_i$$

Because $\widehat{T}_i$ is uncorrelated with u_i, we can now say that the correlation between the treatment variable and the error term is zero in the second stage. In other words, voucher take-up is no longer systematically correlated with the error term. From the 2SLS estimates, Chetty, Hendren, and Katz (2016) find that children who moved to better-off neighborhoods before the age of 13 years had better rates of college attendance, higher earnings, and lower rates of single parenthood as compared to children who did not get the opportunity to move. When applied to a broader context, programs such as MTO are likely to reduce intergenerational transmission of poverty and inequality.

IV is also useful in evaluating infrastructure programs, which are often targeted toward specific areas. In South Africa in 1993, where only one-third of the households had access to electricity, the government committed itself to universal electrification. By 2001, almost a quarter of households were newly connected to the grid due to mass rollout of elec-

tricity. Evaluating the causal effects of the intervention is not straightforward, as program implementation was not random.

To address the selection bias, Dinkelman (2011) uses an IV approach and analyzes the impact of access to grid electricity on employment growth in rural communities. Electrification implementation is instrumented using land gradient. Land gradient is an important determinant of implementation sequence as more time and resources might be required to connect communities residing in higher altitudes and therefore they might be connected to the grid later compared to communities residing on flat lands. The exclusion restriction of the study is that land gradient is unlikely to affect employment outcomes other than through electrification.

Moreover, IV is also suitable to evaluate the effect of good governance on economic growth, which often suffers from endogeneity because governance and economic growth affect each other simultaneously, that is, good governance can increase economic growth but at the same time economic growth can lead to improved governance. Mauro (1995) analyzes data from 70 countries with information on corruption, red tape, and efficiency of the judicial system. Among these institutional factors, he finds that corruption is the cause for lower private investment, which leads to lower economic growth. The IV used to address endogeneity is the index of ethnolinguistic fractionalization, which measures the probability that two persons drawn at random from a country's population will not belong to the same ethnolinguistic group.

The IV meets the two conditions of relevance and exogeneity—countries with higher fractionalization are expected to be more corrupt as bureaucrats may favor their own ethnolinguistic groups; and fractionalization is not expected to directly affect economic growth other than through its effect on institutional efficiency. Not only does the study identify the channel through which governance affects economic growth, but it also estimates the magnitude of the effects, which offers valuable insights into policy-making. For example, the findings suggest that if Bangladesh improves its integrity and efficiency of bureaucracy to the level of Uruguay, its investment rate would rise by almost 5 percentage points and its annual GDP growth rate would rise by over 0.5 percentage points.

Advantages and Limitations of IV

IV enables evaluators to obtain unbiased estimates of treatment effects even in the presence of imperfect compliance. A significant advantage of IV is that evaluators can apply the method even to post-program cross-

sectional data. A drawback is that it is not always feasible to find a valid IV. Unless the IV satisfies the validity conditions, the estimates of the program effect will be biased. Since there is no statistical test for exclusion restriction, one has to draw upon theory and policy background to argue that the IV is truly exogenous. Only under a very specific condition of availability of multiple IVs can a statistical test for weak instruments be conducted.

Propensity Score Matching (PSM)

How can we evaluate a program if we do not have pre- and post-policy data, a clear eligibility criterion, or a valid IV? A quasi-experimental method available to us under such circumstances is propensity score matching (PSM). It can be applied when we only have post-program data. PSM constructs an artificial comparison group by selecting units from the untreated group that share similar observed characteristics with the treated units. As long as there is an untreated group, PSM does not require explicit treatment assignment rules. Another important feature of PSM, that it does not require one-to-one matching of all the relevant observed characteristics, has opened up opportunities for program evaluators to apply matching techniques in IE.

The first step in PSM is to compute the propensity score, which is the probability of being treated calculated using observed characteristics, including factors that influence treatment assignment as well as the outcome. This is done by running a probit or logit regression with the treatment dummy as the outcome variable and all relevant observed characteristics as covariates. The calculated predicted probabilities are then used to identify the treated and untreated units that have the same or extremely close propensity scores. Similar or close propensity scores imply that the treated and untreated units share the same characteristics. The matched treated and untreated units then form the treatment group and the (artificial) control group.

To further explain PSM, let us consider the study by Capuno and Garcia (2010) on the evaluation of a good governance program in the Philippines. The Good Governance and Local Development Project (GGLD) was established with the aim of institutionalizing a set of indicators to track the performance of local governments in the Philippines called the Governance for Local Development Index or Gofordev Index (GI). GGLD was first implemented in 12 local governments in the Bulacan and Davao del Norte

provinces during 2001–2003. In eight out of twelve local governments, GI scores were generated and disseminated to the public to make them aware of the performance of their local governments. In the remaining local governments, the GI scores were generated but not disseminated.

GI assessed the Local Government Units (LGU) based on three performance domains: public service needs (access to and adequacy of basic services and the perceived effectiveness of the LGU in improving family welfare), expenditure prioritization (share of health, education, and other basic services in total fiscal outlays), and participatory development (functioning of the local consultative bodies and the public consultations at the village level). As citizens in LGUs with and without GI dissemination were not directly comparable, PSM was used to generate comparable treatment and control groups. The objective of the evaluation was to examine whether better knowledge of the performance of local government increased civic participation among citizens. The civic participation outcomes are dummies indicating membership in local organizations and participation in local projects.

The propensity scores are computed using a probit regression that controls for all possible relevant observed characteristics that determine the probability of knowledge of GI and also the outcomes. This can be written in the following regression form:

$$P(w_i = 1|\boldsymbol{X}) = G(\boldsymbol{X}\beta) \equiv p(\boldsymbol{X})$$

where w_i is the probability that the individual is in the treated LGU conditional on all the observed characteristics captured in the vector $\boldsymbol{X}$. The propensity score is denoted by $p(\boldsymbol{X})$. In order to minimize selection bias, each individual in LGUs where GI scores were disseminated is matched with an individual in LGUs where the scores were not disseminated. This is done using the computed propensity score $p(\boldsymbol{X})$. The PSM estimates from this study suggest that knowledge of GI led to higher probability of participating in local organizations and civic activities.

Since the treatment effect estimation is done only using the matched units, the resulting estimate is the ATT. Further assumptions required to conduct PSM are common support and unconfoundedness. Common support ensures that treated units have untreated units "nearby" in the propensity score distribution. Common support can be visualized by plotting histograms of treated and untreated units across the propensity score

distribution. The expectation is to see a significant overlap, suggesting a "good match" as shown in Fig. 2.7. The unconfoundedness assumption implies that program participation is determined solely by observed characteristics. This is a strong assumption, and a limitation is that there is no statistical test to prove that there are no unobserved characteristics that affect program participation. However, there are ways to conduct sensitivity analysis to unobserved confounders.

An evaluation of forest protection policies illustrates the use of PSM in assessing environmental policies that suffer from selection bias. Nelson and Chomitz (2011) addressed the fact that protected areas are more concentrated on lands that are unattractive to agriculture, which typically are remote areas with higher slopes and higher elevations because it is easier for governments to implement protection where population density is low and there is less objection (Fig. 2.8).

In such a scenario, an unbiased comparison of deforestation rates between protected and unprotected areas would overestimate the effects

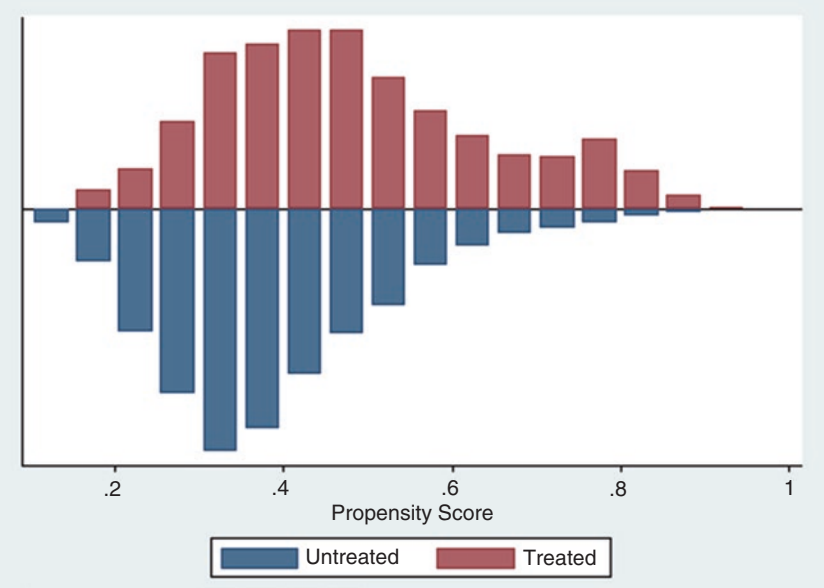

Fig. 2.7 Frequency distribution of treated and untreated units on common support. (Source: Capuno and Garcia 2010)

Fig. 2.8 Protected areas established by 2000. Protected area category: strict (green), multiple use (yellow), indigenous (pink). (Source: Nelson and Chomitz 2011)

of protection. The study used data from developing countries in Latin America, Africa, and Asia and constructed a counterfactual by matching the protected areas and unprotected areas using matching criteria of distance to road network, distance to major cities, elevation and slope, and rainfall. The results showed that the incidence of deforestation is much less in the protected areas than in the unprotected areas.

The study compared the effects of forest protection policy in strictly protected areas, which allow only conservation-related use; multiple-use protected areas, which allow some sustainable use by local inhabitants; and indigenous areas. The general finding was that forest protection policy in multiple-use protected areas was at least as effective as strictly protected areas, suggesting that global environmental goals and local productive activities are compatible. The policy implication derived from this valuable evidence is that setting policies such that there are variations in land use restrictions can be effective in biodiversity conservation and climate change mitigation.

Advantages and Limitations of PSM

PSM is a useful tool to estimate program impact in that it can be applied retrospectively as long as the appropriate data are available. It is desirable to have baseline data, but matching can still be conducted with only post-program cross-sectional data. When there is no baseline data, however, finding all relevant observed covariates is typically a challenge. Further, satisfaction of the common support assumption requires having a large number of treated and untreated units so that a substantial region of common support can be found (usually a large data set). Moreover, as discussed above, unconfoundedness is a strong assumption to make, and

therefore conducting sensitivity analysis to bias from unobserved factors becomes necessary.

Choosing an Impact-Evaluation Method

We have reviewed a number of methods, each of which comes with its own advantages and limitations. How does an evaluator choose which method is best suited to evaluate a particular program? Important questions need to be asked to help determine the most suitable method.

(i) What are the available resources and constraints?
Randomized experiments, by their very nature, are resource and time intensive. Resources needed include financial support and trained man power. A well-designed experiment in a resource-poor environment is bound to fail. Experiments also require pre-intervention or baseline data and a series of post-intervention surveys to be able to capture the treatment effects. While quasi-experiments are less demanding on time and financial support, they still require trained man power to conduct careful econometric analyses. Further, quasi-experiments require good-quality primary or secondary data that are either cross-sectional or panel and have a large sample size, so that the estimates have internal and external validity. Adequate planning and resources are necessary to collect large-scale, nationally representative surveys or panel data.

(ii) Who are eligible units and how are they selected?
Especially in the case of choosing a quasi-experimental method, it is important to know whether there is a well-defined eligibility rule and whether the eligible and non-eligible units complied with the rule.

(iii) What is the nature and stage of the program being evaluated?
In choosing a suitable evaluation method, knowing the scale of the program is helpful. If it is a pilot program or a small-scale intervention, then conducting a randomized experiment might be feasible. In the case of a program that will be nationally rolled out, it may not be feasible to randomize. There are some examples of conducting RCTs at scale, but these require buy-in from policy-makers at the highest level and significant resources (Muralidharan and Niehaus 2017). Therefore, quasi-experimental methods might be

more suitable if appropriate planning and study design is done to collect baseline data. Yet another consideration is the implementation stage of the program. If the program has not commenced, then it might be possible to randomize and collect baseline data. However, if the evaluation is being done ex post, which is mostly the case, then only quasi-experimental methods are suitable.

(iv) What are the outcomes of interest?

A standard way of thinking about outcomes or indicators is that they have to be SMART—specific, measurable, attributable, realistic, and time bound. If the outcomes are not specific or relevant to the objectives of the program, then evaluating them may not be appropriate at all. For quantitative or econometric IEs, it is also necessary that the outcomes are measurable or operationalizable. Further, changes in the outcome need to be attributable to the program to justify conducting an IE. This again emphasizes the relevance of the indicators. Outcomes also need to be realistic in that they are actually achievable through program implementation. In addition, they have to be time bound, that is, evaluators and policy-makers should know when to expect the program to result in the expected outcomes. This may determine whether a quasi-experiment using cross-sectional data is sufficient or whether long-term follow-up, either through experiments through or panel data, is required.

Choosing an appropriate evaluation method by no means necessitates that only one method be used. In fact, combining methods might be a good way to increase the statistical validity of the estimated treatment effects. It is almost a norm to use IV in experiments where compliance is imperfect or where program take-up is driven by self-selection. More and more evaluations are combining methods such as DID and IV or PSM and DID to increase the internal validity and robustness of their estimates. In doing so, it is also important to examine whether the program implementation satisfies the assumptions and conditions of the chosen methods. Table 2.2 summarizes key features of the methods we discussed in this chapter.

Table 2.2 Comparison of key features of empirical evaluation methods

Methodology	*Description*	*Key assumption*	*Data requirements*	*Advantages*	*Limitations*	*Parameter(s) of interest*
Randomized controlled trial (RCT)	Eligible units are randomly assigned to a treatment or control group using an experimental study design	Treatment and control groups are statistically identical in expectation with respect to observed and unobserved characteristics	– Data on treatment variable – Post-treatment outcome data for treatment and control groups – Pre-treatment data on outcome and other characteristics for treatment and control groups to check balance	Generates internally valid impact estimates under the weakest assumptions	– Compliance influences the validity of randomization – Randomization is not always politically feasible – Requires prospective planning and design	Intent to treat (ITT), average treatment effect (ATE), average treatment effect on the treated (ATT), local average treatment effect (LATE)
Instrumental variable (IV)	An IV is used to generate exogenous variation in the endogenous treatment variable	The instrument affects the treatment variable but does not directly affect outcomes (exclusion restriction)	– Data on treatment variable – Data on IV – Post-treatment data on outcome and other characteristics for all units	– Can be applied retrospectively – Estimated causal effects are unbiased even in the presence of imperfect compliance	IV exogeneity cannot be statistically tested except under specific condition of multiple IVs	Local average treatment effect (LATE)

(*continued*)

Table 2.2 (continued)

Methodology	*Description*	*Key assumption*	*Data requirements*	*Advantages*	*Limitations*	*Parameter(s) of interest*
Difference-in-differences (DID)	Estimates the change in outcome over time between treatment and control groups	In the absence of the treatment, changes in outcome for the treatment and control groups are not different (parallel trends)	– Data on treatment variable – Pre- and post-treatment data on outcome and other characteristics for both treatment and control groups	– Controls for entity- and time-invariant unobserved characteristics – Can be used with repeated cross-sections or panel data	Entity- and time-variant unobserved characteristics cannot be controlled for	Average treatment effect (ATE)
Regression discontinuity design (RDD)	Eligible units are determined by a cutoff based on a continuous (running) variable on which the population can be ranked and which is systematically related to the assignment of the treatment	Units that are immediately above the cutoff and immediately below the cutoff are statistically identical	– Running variable and eligibility cutoff to determine treatment status – Post-treatment outcome data and other characteristics for all units	– Does not require any eligible units to be untreated for the purposes of the impact evaluation – Eligibility cutoff can be either sharp or fuzzy	– Treatment effects around the discontinuity may not be generalizable to the entire treatment group – Estimates can be sensitive to inclusion of different functional forms	Local average treatment effect (LATE)

(*continued*)

Table 2.2 (continued)

Methodology	*Description*	*Key assumption*	*Data requirements*	*Advantages*	*Limitations*	*Parameter(s) of interest*
Propensity score matching (PSM)	Creates control group from nonparticipants based on matched observed characteristics of the treatment group	There are no characteristics other than the ones used for matching that affect the treatment status	– Data on treatment variable – Post-treatment outcome data and other characteristics for all units	Can be applied retrospectively	Requires large sample size for valid matching	Average treatment effect on the treated (ATT)

Source: Authors' illustration

Note: Treatment variable refers to policy or program variable whose effect the researcher wishes to evaluate

Challenges in Conducting Impact Evaluations

IEs of programs and policies can be valuable inputs into the assessment of how goals of sustainable development are being planned for and met. Aside from challenges of scope, formulation, and presentation of the key issues, technical, organizational, and political challenges can seriously impede the IE process.

Technical Challenges

Technical capacity includes experts who have skills in data collection, data management, and data analysis. In most developing and less-developed countries, training in social sciences, public policy, and quantitative skills is still lacking. Governments and organizations in these countries often have to rely on aid agencies or external evaluators, who may lack local knowledge. Consequently, methods and indicators used for evaluation may not be suitable to the country context, and the evaluation results may not be useful for decision-making purposes.

Policy-makers might support IEs to gain political credibility, but without trained manpower, this may not be feasible. Overcoming these technical challenges requires building relevant human capital and skills.

Organizational Challenges

Organizational capacity refers to administrative coordination as well as financial resources. IEs are rarely institutionalized, in that they do not follow a systematic approach in identifying, implementing, and using evaluations to inform policy decisions (Bamberger 2009). This requires buy-in and participation from all levels within the organization. This remains a challenge as officials may not view participating in evaluations as part of their responsibilities, especially if tenure and promotion are not linked with achieving program outcomes. Conducting relevant, high-quality, and timely evaluations requires close coordination and alignment of goals among policy-makers, organizations, and the evaluators or technical experts.

A further organizational challenge is budget or financial resources. Integrating IE ex ante in policy design requires committing a significant amount of resources to conduct consultations with various stakeholders, collecting pre- and post-policy data, and conducting and disseminating findings. While this is the ideal case scenario in IEs, resource challenges mean that evaluation is usually done ex post.

Political Challenges

While technical and organizational challenges can be addressed by investing in training and organizational learning, overcoming political constraints can be particularly difficult. Policy outcomes can have significant implications on voter preferences and aid agency assessment. Policymakers may therefore be reluctant to conduct IEs, cherry-pick areas where an evaluation can be conducted, or refuse to accept findings from rigorously and independently conducted evaluations because they do not align with voter expectations.

Organizations may have great interest in assessing whether they have achieved their intended objectives. However, if they directly conflict with political interests, then policies may never be put under the evaluation scanner. These challenges defeat the very purpose of conducting IEs. In extreme situations they can make it impossible to conduct any evaluations.

Conclusions

Evidence-based policy-making calls for the use of findings from IEs, whose scope can vary a great deal depending on the questions asked and the availability of data and other resources. The key value added by IE is in delineating how much of an impact can be attributed or causally linked to a specific policy. While applying IE to understand the effectiveness of policies pertaining to inequality, environmental protection, and governance is thought to be challenging, we demonstrate through real-life policy examples how these tools can be applied to address these big issues.

Often, evaluators are at variance when it comes to "attribution" versus "contribution." IE places clear emphasis on causal attribution. However, when an intervention is complex and involves multiple stakeholders and various aspects, such as economic tools, institutional changes, and social reforms, it might become challenging to attribute changes in outcome to one stakeholder or one aspect alone. At most, evaluators can identify various factors that contribute to the overall outcome.

Contribution analysis can be conducted using logical frameworks and qualitative methods such as in-depth case studies and participatory assessment involving different stakeholders, and it can help to understand what value is added by specific stakeholders or individual components of an intervention to the overall outcome. However, contribution and attribution need not be conflicting objectives of the evaluation exercise. In fact,

contribution analysis can potentially form the basis of future IEs and thus be more complementary to attribution analysis.

It might be farfetched to suggest that one experiment or quasi-experiment can provide all the answers to complex problems that lie at the core of sustainable development. However, cumulative knowledge accumulated through multiple evaluations conducted in multiple contexts will enable policy-makers to provide answers that are rigorously grounded in evidence.

Notes

1. Expectation or expected value refers to the mean of a random variable.
2. See Hahn, Todd, and Van der Klaauw (2001) for details.
3. See Angrist, Imbens, and Rubin (1996) for detailed discussion on the methodology.

Bibliography

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91 (434):444–455.

Bamberger, Michael. 2009. Institutionalizing Impact Evaluation Within the Framework of a Monitoring and Evaluation System. Washington, DC: World Bank.

Cameron, Drew B., Anjini Mishra, and Annette N. Brown. 2016. "The Growth of Impact Evaluation for International Development: How Much Have We Learned?" *Journal of Development Effectiveness* 8 (1):1–21.

Capuno, Joseph J., and M. M. Garcia. 2010. "Can Information about Local Government Performance Induce Civic Participation? Evidence from the Philippines." *Journal of Development Studies* 46 (4):624–643.

Chen, Yvonne, Namrata Chindarkar, and Yun Xiao. 2019. "Effect of Reliable Electricity on Health Facilities, Health Information, and Child and Maternal Health Services Utilization: Evidence from Rural Gujarat, India." *Journal of Health, Population and Nutrition* 38 (7):1–16.

Chen, Yuyu, Avraham Ebenstein, Michael Greenstone, and Hongbin Li. 2013. "Evidence on the Impact of Sustained Exposure to Air Pollution on Life Expectancy from China's Huai River Policy." *Proceedings of the National Academy of Sciences* 110 (32):12936–12941.

Chetty, Raj, Nathaniel Hendren, and Lawrence F. Katz. 2016. "The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment." *American Economic Review* 106 (4):855–902.

Deaton, Angus, and Nancy Cartwright. 2016. "Understanding and Misunderstanding Randomized Controlled Trials." *Social Science & Medicine* 210:2–21.

Dinkelman, Taryn. 2011. "The Effects of Rural Electrification on Employment: New Evidence from South Africa." *American Economic Review* 101 (7):3078–3108. doi: https://doi.org/10.1257/aer.101.7.3078.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2007. "Using Randomization in Development Economics Research: A Toolkit." In *Handbook of Development Economics*, 3895–3962. Amsterdam: North Holland.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica* 69 (1):201–209.

Mauro, Paolo. 1995. "Corruption and growth." *Quarterly journal of economics* 110 (3):681–712.

Meyer, Breed D. 1995. "Natural and Quasi-Experiments in Economics." *Journal of Business & Economic Statistics* 13 (2):151–161.

Muralidharan, Karthik, and Paul Niehaus. 2017. "Experimentation at Scale." *Journal of Economic Perspectives* 31 (4):103–124.

Nelson, Andrew, and Kenneth M. Chomitz. 2011. "Effectiveness of Strict vs. Multiple Use Protected Areas in Reducing Tropical Forest Fires: A Global Analysis using Matching Methods." *PloS ONE* 6 (8):e22722.

Ravallion, Martin. 2018. Can high-inequality developing countries escape absolute poverty? *Center for Global Development Working Paper no. 492.* Washington, DC: Centre for Global Development.

Slough, Tara, and Johannes Urpelainen. 2018. Public Policy Under Limited State Capacity: Evidence from Deforestation Control in the Brazilian Amazon. *Mimeo.*

United Nations. 2016. Global Sustainable Development Report 2016. Chapter 2 "The infrastructure – inequality – resilience nexus". New York: Department of Economic and Social Affairs, United Nations.

White, Howard, and David A. Raitzer. 2017. Impact Evaluation of Development Interventions: A Practical Guide. Manila: ADB.

WHO (World Health Organization). 2014. Access to Modern Energy Services for Health Facilities in Resource-Constrained Settings: A Review of Status, Significance, Challenges and Measurement. Geneva: World Health Organization.