2

What Works and What Doesn't?

Beginning with Adam Smith, economists have long tried to understand why some people and countries are rich while others are desperately poor—typically in the hopes of alleviating poverty. Historically, this work was mostly heavy on theory and light on data. Much of the work introduced in chapter 1, for example, was focused on abstract growth and trade models. By contrast, in recent decades development economists have become decidedly more empirical and more reliant on data to understand what works in practice and what doesn't—typically with a strong microeconomic focus. Today, experiments of different forms are a basic tool in the development economist's kit. Taking a development economics class without learning about experiments and the selection problems they solve is like graduating from medical school without knowing CPR. This chapter will introduce you to randomized control trials (RCTs) and other experimental methods we use to understand the impacts of projects and policies on development outcomes.

ESSENTIALS

- Randomized control trials (RCTs)
- The selection problem
- The reflection problem
- Cost-benefit analysis
- Lab and natural experiments
- Market interlinkages

Ed has allergies. Not the dangerous kind some people get from peanuts or bee stings, but the hay fever kind: sneezing, itchy eyes, congestion, and on bad pollen days, a grueling sinus headache. Fortunately, there is a spray he can shoot up his nose that really helps. He's sure of it. Well, he thinks so. Maybe. Alright, there are days when he uses it and still feels pretty messed up, and other days when he doesn't use it but feels just fine.

The problem is, on spring days when puffballs of pollen float through the air like in a Fellini film, Ed doesn't know what *would* have happened if he *hadn't* sniffed the stuff. Those are the days he almost always uses it. When he forgets to, he can't be sure what would have happened if he *had* taken it.

To complicate matters, once he uses that spray, he acts differently. He feels like he can take on any allergen out there! Students observe Ed bicycling through the Davis countryside with the crops in full bloom. He sneezes from time to time, but that's because he's really putting the sniffer to the test and it isn't supposed to work all the time. Right?

In 2011, international development agencies spent an estimated $US147.74 billion to solve problems far more serious than Ed's allergies.[1] Trying to evaluate whether or not development programs work is a lot like figuring out whether allergy medication works. Development programs are a treatment, and the problems they try to solve are like an allergic reaction to pollen.

Donors must have better ways of knowing whether their programs work than Ed has for nose spray, right?

Sadly, until fairly recently they did not. Development agencies' shelves and hard disks are filled with final reports concluding that the projects they funded were successful (usually) at achieving their stated goals. But it can be extremely difficult to show whether a treatment is successful or unsuccessful. That is, unless you've got an experiment.

The people who make nose spray know all about experiments. That's what drug trials are all about. Before they can market a new drug, they have to perform a *randomized control trial,* or RCT. The formula to do a RCT is simple: (1) devise a treatment; (2) identify your target population and from it randomly select a sample of people to run your experiment on; (3) split the sample randomly into two groups, a treatment group and a control group; (4) give the treatment group the treatment and the control group a "placebo" that looks like the treatment but isn't; (5) after enough time has elapsed for the treatment to take effect, gather new information on your treatment and control groups; and (6) compare outcomes of interest between the treatment and control groups.

In 1997, Mexico did something similar to a drug experiment, but it was to test an entirely different sort of treatment: a new welfare program. PROGRESA (Programa de Educación, Salud, y Alimentación) was designed to combat rural poverty from two angles. First, it gave cash to poor people. A number of studies have shown that women are

more likely than men to spend income on food and other goods that benefit their families, so women were the target of the program. Second, in order to get the cash, a poor woman had to follow some rules to improve her family's nutrition, health, and education. Kids had to be enrolled in school and in the local medical clinic. These behavioral requirements made PROGRESA what is called a "conditional cash-transfer program," or "CCT."

The theory behind this CCT was simple. In the short run, cash is what poor people need most in order to feed and clothe their families and satisfy their basic needs and wants. In the long run, the best way to break the intergenerational transfer of poverty is to give kids the human capital they need to lead productive lives; hence the two *C*'s.

So far, we've got most of the first two elements of an RCT: the treatment (the CCT) and a target population (poor rural women). Mexico had to find out who was in this target population, so it carried out a nationwide survey. It identified 2.6 million families in fifty thousand rural communities who were eligible to receive PROGRESA benefits. That's about 40% of all rural families. The plan was to give the PROGRESA treatment to all eligible women.

If all eligible women get the treatment, how can we test whether the treatment works? We could compare everyone before and after the program starts. But if we saw differences, say, in school attendance or family nutrition, could we be sure it was because of PROGRESA? Many other things were happening in Mexico at the same time as PROGRESA. NAFTA (the North American Free Trade Agreement) had just gone into effect. In Mexico, as in many other countries, the mid-1990s saw far-reaching agricultural reforms that included eliminating subsidies for small farmers, with big impacts on rural incomes. New rural schools were being built. People were migrating. The weather was changing. There was lots of pollen in the air.

If you give everyone an allergy spray, they might still sneeze if the pollen count rises—or they might not sneeze at all if it doesn't. When you can't control for everything else, you can't figure out whether your treatment worked. Something else might have changed. This has been the curse of development-program evaluations over the years. We need a control group of similar, randomly chosen people who did *not* get the treatment but experienced, on average, the same changes in all those other variables that the treated people did. If treatment and control groups go into the pollen together, we should be able to determine whether the drug works.

Fortunately, the way PROGRESA was rolled out created a random control group for evaluating the program's impacts. There was no way to roll out the program to all eligible families in rural Mexico at the same time, so the government had to choose which poor villages to "treat" first. It could have gone for the villages closest to Mexico City, near where powerful politicians lived, or where poverty was highest, but it didn't. Instead, it rolled out the program randomly. All eligible women in randomly chosen villages got PROGRESA payments the first year of the program. They were the treatment group. In the rest of the villages, none of the eligible women got PROGRESA right away. They were the control group.

Randomization ensured that the treatment and control villages, households, and women, on average, were identical except for the treatment, just like the treatment and control groups in a drug trial. Researchers could compare any outcome they wanted—school attendance, nutrition, whatever—between the eligible households in these two groups of villages. All you had to do was compare averages. The difference could be attributed to PROGRESA.

Within three years, all 2.6 million eligible families were getting PROGRESA, so the experiment vanished. But for a short period of time, Mexico had given the world the gift of a randomized "social experiment" in the form of an RCT (see sidebar 2.1). PROGRESA became the model for both designing and evaluating anti-poverty programs in many other developing countries and even in New York City.[2]

RANDOMIZATION AND THE SELECTION PROBLEM

Over the years we've noticed that people who use nose sprays sneeze more than people who don't. Could it be that nose spray *makes* you sneeze?

That's a silly question, you say. People who use nose spray sneeze more because they had more allergies to begin with; that's why they chose the nose spray treatment. That's probably true, but you can see the problem here. We cannot determine whether the nose spray is effective by comparing people who use it with people who don't. If we do that, we might well conclude that the drug makes people sneeze! This is what experimentalists call *selection bias*. Selection bias confounds all sorts of studies. Here are three illustrations:

The economists Joshua Angrist and Jörn-Steffen Pischke took people who were hospitalized (the treatment group) and people who were not

Sidebar 2.1 Progressing with PROGRESA

Mexico's PROGRESA data have spawned more development economics research (not to mention PhD student theses) than almost any other micro data set in the world. Here are some key findings on PROGRESA's impacts, all made possible by the way in which the program was randomly implemented across rural Mexico.

Nutrition: PROGRESA improved both calorie consumption and the quality of beneficiaries' diets. Eligible households in treatment localities consumed 6.4% more calories than comparable households in the control localities. When it comes to nutrition, the quality of calories also matters. The study found that PROGRESA's biggest impact was on calories from vegetable and animal products. PROGRESA made people eat not only more, but better.

J. Hoddinott and E. Skoufias, "The Impact of PROGRESA on Food Consumption," *Economic Development and Cultural Change* (October 2004):37–61.

Schooling: PROGRESA had a significant positive effect on school enrollment. Many kids drop out of school after grade 6, when often they must leave their village to continue on in school. The largest difference between PROGRESA and control households was for kids who had already completed grade 6; the PROGRESA kids' enrollment rate was 11.1% higher, reaching 69%, and the program's impact was disproportionately concentrated among girls. Exposure to PROGRESA for 8 years, starting at age 6, increases children's educational attainment by an average of 0.7 years, and 21% more children attend secondary school.

T. Paul Schultz, "School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program," *Journal of Development Economics* 74, no. 2 (2004):199–250.

Jere R. Behrman, Piyali Sengupta, and Petra Todd, "Progressing through PROGRESA: An Impact Assessment of a School Subsidy Experiment," *Economic Development and Cultural Change* 54, no. 1 (2005):237–75.

Health: PROGRESA significantly increased preventive care, including prenatal care, child nutrition monitoring, and adult checkups. It reduced inpatient hospitalizations, suggesting a positive effect on major illness. PROGRESA children age 0–5 had a 12% lower incidence of illness, and prime age adults (18–50) had 19% fewer days of difficulty due to illness than did non-PROGRESA individuals.

Paul Guertler, "Final Report: The Impact of PROGRESA on Health" (Washington, DC: International Food Policy Research Institute, 2002 (www.ifpri.org/sites/default/files/publications/gertler_health.pdf)

(the control group) and compared their health status a year later.[3] The people who had been hospitalized were less healthy. Do hospitals make people sick?

Governments around the world offer job training programs. Many studies find that a year or two later the people who chose to be in these programs are more likely to be employed than the people who chose not to do the job training. Are job training programs successful, or is it the kind of person who chooses to go for job training?

Economic studies consistently show that people with more education have higher earnings. Is this because schools make people more productive, or do higher ability people go to school?

In these (and countless other) cases, the outcomes we see after the treatment reflect two things: first, who chooses to get the treatment (the selection effect), and second, the effect of the treatment, itself. Because of this, simply comparing outcomes for people who did and did not get a treatment may tell us nothing at all about whether the treatment was effective. We've got to untangle the two.

What we'd really like to do is compare the same person's outcome with and without the treatment. We can't do that, though, because once a person gets treated, we can't see what would have happened to her without the treatment. And if the person does not get the treatment, we'll never know what would have happened if she had been treated.

The selection problem arises when things that determine whether or not someone gets treated are correlated with the outcome we want to measure. Sick people (whether they go to hospital or not) are likely to be less healthy in the future. Motivated people choose to participate in a training program, but they are more likely to get a job with or without the program. High-ability people are more likely to have higher earnings, regardless of how much more productive schools make them.

Randomization solves the selection problem. By randomly choosing who gets the treatment and who does not, RCTs create treatment and control groups that on average are the same except for the treatment. Any differences we observe between the two, then, must be the result of the treatment. You can find a formal presentation of the selection problem and how randomization solves it in the appendix to this chapter, "The Math of Selection."

Theoretically, in a perfectly designed experiment, we could test whether or not the treatment is successful simply by comparing outcomes between treatment and control groups. Randomization would

ensure that everything but the treatment is identical, on average, between the two groups. Real life rarely gives us something approaching perfect randomization, though. Thus, we usually need baseline (pretreatment) information to make sure the treatment and control groups really are the same except for the treatment. Baseline surveys are costly, but tests showing there are no significant differences between the treatment and control group prior to the treatment are important to validate RCTs.

Baseline surveys are important for other reasons. We saw previously that Mexico's PROGRESA had to carry out a baseline survey in order to find out who would be in its target population, that is, which women met the criteria for receiving PROGRESA payments.

Baseline information can help researchers control for other variables that affect the outcome of interest. For example, while treatments are carried out, other things in the economy are changing, like the weather, macroeconomic policies, and recessions. With good baseline data, we can compare *changes* in outcomes for the treated and control groups before and after the treatment. For example, we might hope that cash transfers raise crop production in poor households. Meanwhile, if the economy is growing, poor households might increase their crop production with or without the program. If the transfers really do increase crop production, though, *the change in crop production should be larger in the households that got transfers*. Instead of comparing crop production between treated and nontreated households, then, we can learn more about the program's impacts if we compare *differences* in crop production between the two groups. This is called the "difference in difference" method. We first calculate the difference in the outcome variable (crop production) before and after the treatment for both the treatment and control groups. Then we calculate the difference between these differences. If it's positive, we conclude that the treatment had a positive effect on the outcome. This useful method requires having data on the treated and control groups before as well as after the treatment.

THE EXPERIMENTAL REVOLUTION IN DEVELOPMENT ECONOMICS

The chief architect behind PROGRESA was an economist named Santiago Levy who got his PhD from Boston University in 1980. By the time Mexican president Ernesto Zedillo (an economist with a PhD from Yale in 1974) asked him to lead a team to address extreme poverty in Mexico, Santiago had done enough data analysis to appreciate how

selection bias can make it tough to know whether any program actually worked in practice.

The program he and his team launched was the first large-scale randomized policy experiment in a developing country. The RCT approach to evaluating its impact was inspired by the work of economists studying policies in developed countries (especially related to labor markets). This, in turn, set the stage for a revolution in how development economists try to learn what works and what doesn't. The essence of this methodological revolution is quite simple: the less choice people have about whether to be "treated," the easier it is to test what works and what doesn't.

Many development economists see RCTs as the impact-evaluation gold standard, because in their purest form RCTs do not permit people to have any choice about whether or not they are treated. In 2003, Esther Duflo cofounded the Poverty Action Lab, which is dedicated to the use of RCTs.[4] She writes: "Creating a culture in which rigorous randomized evaluations are promoted, encouraged, and financed has the potential to revolutionize social policy during the 21st century, just as randomized trials revolutionized medicine during the 20th."[5] The J-PAL website states: "Randomized evaluations are often deemed the gold standard of impact evaluation, because they consistently produce the most accurate results . . . to determine whether a program has an impact, and more specifically, to quantify how large that impact is."[6]

Today, RCTs are being used to evaluate a wide array of development programs, from a new generation of social cash transfer (SCT) programs in sub-Saharan Africa to microcredit, HIV/AIDS prevention, immunization, and even "hope." Here are a few examples of the kinds of questions RCTs address.

RCTs for African SCTs

African countries are different from Mexico in ways that could shape the outcome of cash transfer programs. They are poorer and characterized by a greater level of risk and vulnerability. African SCT programs typically target households that are labor-poor as well as being in extreme poverty and containing vulnerable children. HIV/AIDS has its global epicenter in Southern Africa. The region has less developed markets and greater political instability. People's livelihoods and ability to escape from poverty are more linked to small-holder agriculture and the informal economy than to the formal wage economy. Public institutions

tend to be weaker, and governments have fewer resources to invest in poverty programs, and thus international donors play a much more significant role in financing social programs in sub-Saharan Africa. Competing donors often have conflicting ideas as to the types of social protection interventions to pursue. There is a lack of consensus among governments, too, along with a weaker capacity to implement and evaluate programs, and fewer complementary services like health, education, and nutrition. All these considerations make sub-Saharan Africa both an important laboratory for impact evaluation and a challenging place to do it.[7]

Another fundamental difference between the African and Mexican programs is that, for the most part, the African programs are not conditional. Often, behavioral changes like better nutritional practices and keeping kids in school are encouraged, but with few exceptions they are not required as a condition of getting the transfer. These programs are often referred to as social cash transfer (SCT) instead of CCT programs. Is conditionality really needed, or, given the cash and information, will people choose to do the right thing? These questions loom in the debate and evaluation of SCTs in sub-Saharan Africa. There are exceptions. Ethiopia's Productive Safety Net Program (PSNP) pays people from eligible households in chronically food-insecure *woredas* (districts) to work on labor-intensive projects. It is conditional in the sense that people have to work in order to get benefits. The idea behind this project is to give cash and food to the poor while building up the country's infrastructure, particularly irrigation, via work projects in which the beneficiaries participate.

A number of evaluations have come out of pilot programs designed to test the effectiveness of SCTs before the programs are "scaled up" to the larger population. Sidebar 2.2 summarizes what some of the key African SCT evaluations have been finding. As we'll see in the next section, randomizing the "SCT treatment" is the key to being able to make statements like these about causality.

Credit

Access to credit is vital to people in poor as well as rich countries, as we shall see in chapter 12. There is strong theoretical reason to think that people will invest in new activities and technologies when they get access to credit. But how big is the impact? Do microcredit projects really make people more productive, and if so, how much?

Sidebar 2.2 Impacts of SCTs in Sub-Saharan Africa

An evaluation of a pilot SCT program in Malawi showed a significant reduction in child morbidity, gains in school enrolment, and increases in food consumption and diet diversity. Agricultural investments increased. The SCT also reduced child labor outside the home.

C. Miller, M. Tsoka, and K. Reichert, "Impacts on Children of Cash Transfers in Malawi," in *Social Protection for Africa's Children,* edited by S. Handa, S. Devereux, and D. Webb (London: Routledge, 2011), 96–116.

Katia Covarrubias, Benjamin Davis, and Paul Winters, "From Protection to Production: Productive Impacts of the Malawi Social Cash Transfer Scheme," *Journal of Development Effectiveness* 4, no. 1 (2012):50–77.

Ethiopia's Productive Safety Net Program caused an increase in school attendance for some groups, particularly younger children, and a reduction in child labor for some activities among boys, but an increase in girls' labor time.

J. Hoddinott, D. O. Gilligan, and A. S. Taffesse, "The Impact of Ethiopia's Productive Safety Net Program on Schooling and Child Labor," *Social Protection for Africa's Children,* edited by S. Handa, S. Devereux, and D. Webb (London: Routledge, 2011), 71–95.

South Africa's Child Support Grant decreased school absences, illnesses, and hunger and increased height-for-age scores among children receiving the grant. It increased access to cell phone use and supported the sustainability of agricultural activities in households with children receiving the grant. It also significantly reduced risky behaviors among adolescents, including sexual activity, pregnancy, alcohol use, drug use, criminal activity, and gang membership.

DSD, SASSA, and UNICEF, *The South African Child Support Grant Impact Assessment: Evidence from a Survey of Children, Adolescents and Their Households* (Pretoria: UNICEF South Africa, 2012; www.unicef.org/evaldatabase/files/CSG_QUANTITATIVE_STUDY_FULL_REPORT_2012.pdf).

Kenya's Cash Transfers for Orphans and Vulnerable Children (CT-OVC) increased children's secondary enrollment on par with what has been found from *conditional* cash transfer programs in other parts of the world. Participating households had significantly higher expenditures than control households in food, health, and clothing and significantly less spending on alcohol and tobacco. They shifted from tubers to cereals, meat and fish, and dairy.

The Kenya CT-OVC Evaluation Team, "The Impact of Kenya's Cash Transfer for Orphans and Vulnerable Children on Human Capital," *Journal of Development Effectiveness* 4, no. 1 (2012):38–49.

Testing the effect of credit on investments and other outcomes is difficult, because the kinds of people who get loans (i.e., they apply and are accepted) are different from the kinds that do not, so we cannot simply compare the two. How can you make an experiment out of credit?

Dean Karlan and Jonathan Zinman figured out a way.[8] They convinced a lender in South Africa to grant loans to a random sample of applicants with low credit scores. These were people who applied for credit but had been deemed not credit worthy. Giving credit to people who do not qualify for it might not seem like the best idea, and it raises some ethical concerns (see "The Ethics of Experiments" later on in this chapter), but by randomly giving credit to people in this group, Karlan and Zinman avoided the problem that more credit-worthy people get loans, and they might do well with or without credit. It was an RCT because only some randomly chosen people with low credit scores were given loans, while others were not.

The researchers compared those who got credit to those who did not in terms of "economic self-sufficiency" (employment and income), food consumption, and other outcomes six to twelve months after the treatment. They found that economic self-sufficiency and food consumption were higher for the treated group. They also found that depression and stress were higher for the people who won the loan lottery, perhaps due to anxiety from being in debt—a result that raises an obvious ethical concern about giving credit to borrowers who are not credit worthy.

People didn't randomly incur debt in the RCT that Suresh de Mel, David McKenzie, and Christopher Woodruff did in Sri Lanka—they just got money or machines.[9] The entrepreneurs who got chosen for this "Santa Claus treatment" ended up with a significantly larger capital stock, which is not so surprising for the ones that got the machines but not predictable for the ones that got the cash. However, the effect of this treatment on the profitability of enterprises was small or insignificant. These results suggest that some businesses are constrained by a lack of capital while others are not.

Insurance

Evaluating how insurance affects poor households is challenging because almost no rural households have access to insurance, and those that do have insurance tend to be very different from those that do not. Characteristics of households that are correlated with whether or not

they have insurance are also likely to explain outcomes like crop production, income, or nutrition. Because of this selection problem, comparing outcomes between households that get insurance and those that do not generally tells us little.

We know from past research (see chapter 12) that poor households diversify their activities more than rich households to protect themselves against uncertainty. By not "putting all their eggs in one basket," though, they forfeit the potential income gains from specializing in what they do best. Access to insurance could bring substantial economic benefits to rural households, because if harvests are insured, banks might be more willing to lend to farmers, and farmers might be better able to specialize. To test this, though, we need a treatment group of households that have access to insurance and a control group that does not. We also need to avoid the problems of *adverse selection* and *moral hazard,* which we'll learn about in chapter 12; otherwise, insurance companies will not be willing to offer insurance to small farmers. Where can we find all of this?

Sarah Janzen and Michael Carter came up with a way.[10] They offered a new kind of insurance to a random group of pastoralists in the Marsabit District of northern Kenya: index-based livestock insurance (IBLI). Satellite measures of vegetative cover are used to predict average livestock mortality from drought in local communities. The payout households get from this insurance has nothing to do with their behavior; this insurance pays if the average livestock mortality predicted from satellite images reaches 15%. This avoids the problem of moral hazard (people changing their behavior once they have insurance). Janzen and Carter convinced an insurance company in Kenya to make this insurance randomly available to some small farmers but not others. This helped solve the problem of adverse selection (higher risk people taking out insurance).

A drought hit in 2011, after the insurance was made available. Insured households got an average payout of $150. It is too soon to assess the impact of this insurance, but we can ask what people *think* it will be. Janzen and Carter asked both the insured and uninsured households how they plan to deal with the drought. Many responded that they'll eat fewer meals, but a significantly smaller percentage of those with insurance said this. The number who anticipated selling additional livestock to cope with the drought was 50% lower for the insured households. The insured households also said they will rely less on food aid and assistance from others.

If what households end up doing is anything like what they say they'll do, this project will have succeeded in helping pastoralists deal with drought risk and avoid some of the worst impacts of the drought, while demonstrating the importance of insurance in risky environments.

Hope and Optimism

Most people care about their future. But what if, when they look there, what they see are dim economic prospects? Psychologists call the uncomfortable tension people feel from simultaneously holding conflicting thoughts "cognitive dissonance." Could it be that the poor, by closing their eyes on the future, reduce their psychological distress at the cost of worsening their future economic well-being? If poor people close their eyes on the future, they will have no reason to save and invest for it. This can create a "psychological poverty trap."

In November–December 2010, a team of researchers in Mozambique ran a lottery in which the winners got a free input subsidy for 70% of the cost of a seed and fertilizer package.[11] Winners of this lottery could expect to get a larger harvest. In April–May 2011 both the winners and losers of the lottery were asked the question, "How much time ahead do you plan your future expenditures?" On average, winning the lottery increased an individual's time horizon by more than a month, from 198 days to 235 days. It seems that the farmers who won the lottery became more forward looking.

Another RCT, in India, found evidence that helping desperately poor people invest gave far better results than expected, consistent with breaking out of a psychological poverty trap (see sidebar 2.3).

You will find many other RCTs scattered throughout this book. They test the impacts of a wide variety of programs, from immunizations to HIV/AIDS and government corruption.

RCTS AND THE PRACTICE OF DEVELOPMENT

The introduction to this chapter alluded to a methodological progression in development economics. The core questions about why some people and societies are poor and what might help alleviate this poverty have stayed essentially the same, but the methods economists use to try to address these questions have changed markedly. The approach of Adam Smith and his contemporaries was largely qualitative and even philosophical. As economics became more formalized and mathematical,

Sidebar 2.3 Hope

In the Indian state of West Bengal, a microfinance institution, Bandhan, tried something different. Instead of giving loans to extremely poor people, who they thought would be unlikely to repay, they gave out assets: a few chickens, a cow, a pair of goats. They also taught people in this treatment group how to take care of their animals and manage their households. Just to make sure they wouldn't eat the animals right away, they also gave them a little cash to spend.

The theory behind this RCT was that people would learn how to manage their finances better and make a little income selling the products their farm animals would provide. To test the results of this project, researchers compared these treated households with a random control group of poor households, which did not get any of these things.

The results? The treatment worked better than anyone had hoped for. Long after the treatment had ended, the treated households ate 15% more, earned 20% more, and skipped meals less often than households in the control group. They were saving more, too. The improvements were far too big to be explained by the direct effects of the grants. That is, the treated households could not have sold enough eggs, milk, or meat to explain these big outcomes.

The project gave the treated households more than it had expected. The research team, headed by economist Esther Duflo, called it hope. The project gave people a reason to work harder—28% more hours, to be precise. The incidence of depression fell. In addition to a few animals and a bit of advice, it seemed, BRAC had succeeded in administering a healthy dose of optimism. Could it be that the hope for escaping from poverty traps is hope, itself?

"Hope Springs a Trap: An Absence of Optimism Plays a Large Role in Keeping People Trapped in Poverty," *Economist* (May 12, 2012; www.economist.com/node/21554506).

development economists focused mostly on theoretical models to shed light on these questions. When international development assistance expanded rapidly after World War II, there was a substantial rift between the abstract modeling of ivory tower economists and the emerging ranks of "boots on the ground" development practitioners who designed and managed development projects. These were two different worlds that, at best, shared only poverty questions as a raison d'être.

The availability of data and computing power in the 1970s and 1980s shifted many development economists away from purely theoretical models toward serious empirical analysis—and enticed them to spend more time in the field collecting data. The 1990s brought additional empirical advances to development economics. During these decades development economists tended to interact more and more with development practitioners, but there remained a persistent gap between applied development research and the practice of development.

The experimental revolution of the 2000s has shrunk this gap noticeably and increasingly brought practitioners and economists together as collaborators. In a typical collaboration, an implementing practitioner organization (e.g., an NGO, agency, or company) that intends to launch or expand a project relies on a team of research economists to design an RCT to rigorously evaluate the intervention and conduct the analysis. Success demands careful coordination and close interaction between practitioners and development economists.

These new models of collaboration have shaped both development economics and the practice of development. Greater integration of the two has brought development economists into earlier stages of project design and evaluation. Aligning research and programmatic objectives often also leads economists to analyze more directly the costs and benefits of specific development programs and interventions. (We will learn about how to do a cost-benefit analysis later on in this chapter.) Cost-benefit analysis can build directly on RCTs, which can help quantify impacts and associated benefits.[12] Close collaboration with the implementing partner makes it easier to incorporate program costs into this analysis and determine whether the carefully measured benefits justify the costs required to reap these benefits. Ideally, more rigorous cost-benefit analysis of this sort improves development policy and programs by distinguishing good projects from bad ones.

RCT-based research has also shaped the practice of development in important ways. Development organizations of all sorts now feel pressure to rigorously evaluate their projects. While this presents new opportunities to learn what works and what doesn't, it also brings new risks. The prospect of establishing very clearly and cleanly that something works sounds great, but establishing with equal clarity the opposite looks and feels like failure. This can be a threat, perceived or real, to both reputation and future funding. In contrast, less rigorous impact evaluation methods can be more forgiving, allow organizations to use selection bias to their advantage, and leave plenty of room for casting

evidence in a more favorable light (glossy annual reports often do exactly this).[13]

WHEN EXPERIMENTS CAN GO AWRY

For all their promise, RCTs have many potential pitfalls. We have highlighted a collection of well-executed experiments, but in practice, the ideal experiment is exceedingly hard to find. In general, the best experiments are those in which the question asked lends itself neatly to experimental methods, and the researchers have control over how the experiment is designed and executed. This usually is not the case with large-scale government programs, in which many things can go wrong, from politics to poor administration of treatments and research. Anyone reading about or designing RCTs had better be aware of these pitfalls.

There are two types of pitfalls that are worth noting: technical pitfalls that may undermine what we are able to learn from an RCT, and ethical pitfalls that arise in experimenting with people. We focus initially on the technical pitfalls of RCTs and discuss the ethical considerations, of RCTs in particular and experiments more broadly, later in the chapter.

The following subsections describe a few of the technical pitfalls that may beset an RCT.

Creating Treatment and Control Groups

An RCT requires treating one randomly selected group and denying treatment to another. Before we can conduct an experiment, we need to have valid treatment and control groups. Creating treatment groups is not as easy as it may at first seem. Why not? Because it is often difficult to ensure that the people who are randomly assigned to treatment are actually treated. For example, if you wanted to test the impact of a new crop variety on household income, you could offer incentives to adopt the new variety to the farmers in the treatment group, but you cannot ensure complete compliance. That is, farmers still must choose to plant the new variety—and this choice threatens to introduce the selection bias described above. As we will discuss later in this chapter, economists have devised ways for careful experimental design and data analysis to remedy this potential pitfall.

There are yet deeper potential problems surrounding the creation of control groups, which sometimes simply may not be possible. Take tourism, for example. Ecotourism development projects are among the fastest-growing parts of development bank loan portfolios. Many countries see tourism as a way to stimulate economic growth and fight poverty. Suppose we are interested in quantifying the impacts of a tourism-development project. The treatment is the project. The treatment group is effectively the entire population at the tourist destination, and the control group is the same population without the project. It is not possible to make this project happen for one group of people but not for others at the tourist destination. One might argue that the project could be implemented at some randomly chosen tourist sites but not others. However, almost by definition tourist destinations are unique (hence the reasons tourists want to go there). This makes it difficult to come up with reasonable alternative locations as a control, that is, sites identical to the "treated" site except without the treatment. They simply do not represent the region without the project. There is no counterfactual for the Galápagos Islands.

There are many other cases in which problems arise in the construction of treatment and control groups. Irrigation and other infrastructure projects create public goods that potentially affect everyone in the zone in which the projects are carried out. Staple price supports frequently have been used as a mechanism to transfer income to farmers (with dubious welfare benefits). However, it is generally not feasible to offer a high price to some randomly selected farmers but not others.

It may be politically infeasible to randomly create a treatment group, or it may be considered unethical to deny benefits to a control group (see "The Ethics of Experiments," below). In theory, input subsidies could be implemented randomly through targeted vouchers. In practice, though, it may not be politically feasible to deny benefits to a control group while offering them to a treatment group.[14] Even in a country like Malawi, where fertilizer vouchers targeted poor farmers, they were not given out randomly. If subsidies are given to all qualifying farmers, there is no control group.[15] It is not uncommon for researchers to be called upon to conduct impact evaluations after a project has already been implemented. In this case, we can see who got the treatment and who did not, but we might not have the pretreatment data we need to do a clean RCT, and there might be concerns over whether the creation of the treatment and control groups was truly random.

Control Group Contamination

Measuring a project's impact on the treated requires isolating the control group from the project's effects. This is often not so easy to do in practice. Even in a medical experiment it may be difficult to isolate the control group from the treatment group, for example, if the treatment involves curing a communicable disease. The effects of treatments on control groups frequently confound experimental research in the social sciences.

A well-known RCT in Kenya illustrates this point. It was designed to treat school children with worms in an effort to keep them in school. But by treating kids in some schools, the incidence of worms among kids in *control* schools went down (see sidebar 2.4). When the treatment affects the control group as well as the treatment group, it can be difficult or impossible to reliably estimate the impact of the treatment, because both groups change. We call this problem "control group contamination."

Economic linkages can transmit impacts from treatment groups to others inside and outside the local economy. Take a cash transfer program. The household that gets the cash spends it. In the process, it transmits the impacts of the program to others inside and outside the village. Ed spoke with a shopkeeper in an Ethiopian village who loved the cash transfer program there. "You get transfers?" Ed asked. "No, but the people who get money come here to spend it!" he answered.

This shopkeeper was not eligible for the treatment, but he benefited from it just the same. If treated households buy more food, local farmers can benefit. If they fix up their house, so can the local bricklayer. These people, in turn, may hire more workers and buy more inputs. This can lead to a village version of Keynesian economics, in which the infusion of new cash into the economy has a multiplier effect on village income.[16] If we only look at the treated households, we are likely to underestimate the overall effects of the treatment.

Economic spillovers do not necessarily result in control group contamination, but they may. If the control households are in another village, economic linkages from the treated villages might not reach them. However, all around Africa, periodic markets bring people together from many different villages to buy and sell. If households from treated and control villages interact in these markets, the result can be control group contamination.

Treatment spillover effects raise challenges for RCTs, and they can be good or bad for people. If the treatment positively affects the control

Sidebar 2.4 Worms

Worms are bad (unless they're the garden variety). Hookworm and roundworm each infect approximately 1.3 billion people around the world; whipworm affects 900 million, and 200 million are infected with schistosomiasi. Intense worm infections keep kids from going to school and reduce their educational achievement. Could it be that a key to literacy is (getting rid of) worms?

Edward Miguel and Michael Kremer analyzed a RCT experiment to raise school attendance in Kenya by treating children for worms. A clearly defined treatment for worms was administered to children in a randomly selected sample of schools (the treatment group) but not in other schools (the control group). This project had a simple and easily measured outcome: school attendance. The ex-post research question was whether or not children in the treated schools were more likely to attend school after the treatment.

It seemed to be a squeaky clean experimental design. What could go wrong with it?

Actually, something went too right, from an analytical point of view. The treated schools treated the control schools. Maybe treated kids played with control kids after school or had contact with others who, in turn, had contact with control kids. The study could not tell us why, but for whatever reason, kids in the control schools got better, too.

Miguel and Kremer call this an *externality* of the treatment. (We'll learn about externalities in chapters 6 and 11.) In experimental jargon, it is called control group contamination. Really, it is a linkage—in this case, an epidemiological one—that transmitted the benefits of the project from those directly affected (the kids in the treatment school) to others in the project's zone of influence. Not surprisingly, the authors found that the farther a treated school was from a control school, the bigger the measured impact of the treatment.

Since kids in control schools got better, it was hard to find a positive effect on school attendance by comparing the treatment and control groups. It is ironic that a treatment potentially can be so successful that you cannot show it has any effect at all.

Edward Miguel and Michael Kremer, "Worms: Identifying Impacts on Education and Heal th in the Presence of Treatment Externalities," *Econometrica* 72, no. 1 (January 2004):159–217.

group, we might conclude that the treatment was not effective when in fact it was—both the treatment and control group benefit from it. It is also possible that the spillover is negative. For example, some villagers complain that cash transfers push up food prices. Giving cash to people might lead them to work less, in which case wages could go up. This creates a cost for those who hire workers. If a project negatively affects the control group, we run the risk of concluding that the treatment was effective when really it was not: the treatment appears to make the treated better off when really it makes the control group worse off.

Under ideal circumstances, randomization can ensure that the expected outcome for the control households equals the expected outcome of the treated households had they not gotten the treatment. This ideal randomization relies in fact on two conditions. The first is having a "clean" control group that is isolated from the treatment. That is, it must be absolutely unaffected by the presence or absence of the treatment. The second is that the control group needs to be so similar to the treatment group on average that, had there been no treatment, the two groups would have displayed the same outcome.

Can Development Be Studied like a Pill?

There is no question that the widespread use of RCTs in international development that started with PROGRESA has profoundly changed the way economists, NGOs, aid agencies, and governments approach development problems. RCTs are a major—perhaps the major—focus of development economics today. It is hard to find a development student PhD thesis that does not include some kind of randomized treatment. The strongest proponents of RCTs argue that randomized evaluations are "the gold standard of impact evaluation, because they consistently produce the most accurate results . . . to determine whether a program has an impact, and more specifically, to quantify *how large* that impact is."[17] A big lesson from RCTs is that there is no single solution or explanation for underdevelopment. Different kinds of action are needed in different settings.

Others, we have seen, question whether RCTs are the end-all tool they claim to be and whether the most pressing development questions can be answered using a randomized experiment. For example, one of the most ambitious development interventions of recent decades—the Millennium Village Project—is a grand social experiment, but it was not designed as an RCT. In the minds of many development econo-

mists, this limits our ability to evaluate the project's impact. Jeffrey Sachs—the architect of this project—dismisses these concerns, saying, "Millennium Villages don't advance the way that one tests a new pill."[18] His view, which is shared by some other researchers, is that restricting ourselves to a single methodological approach will severely hamper our ability to understand the complexities of the development process. What do you think?

IF NOT RCTS, THEN WHAT?

Although RCTs have become its poster child, the experimental revolution in development economics extends beyond this method. Recall that the essence of this methodological revolution is that the less choice people have to opt in or out of a treatment, the easier it is to test the treatment effect. RCTs offer a clean and direct way to introduce random treatment (albeit not as cleanly and simply as they might seem, as discussed above), but there are experimental alternatives, specifically (1) laboratory experiments and (2) natural experiments. Let's look at each of these in turn.

Laboratory Experiments

In a laboratory experiment, subjects make economic decisions in a contrived setting that is designed by researchers to elicit a specific kind of response from them. Travis has designed and conducted economic experiments like this in India, Morocco, Bolivia, and in several sub-Saharan African countries to learn how people make decisions. Individuals' responses in these controlled settings help us measure things that are otherwise difficult to measure: aversion to risk, patience, trust, concerns about fairness, and willingness to cooperate, to name just a few.

These "laboratory experiments" rarely take place in a laboratory when development economists do them. Instead, trained teams (often consisting of local university students or recent graduates) conduct these experiments in places where people tend to gather: under trees, in health clinics, near village schools, and so on. The structure of these experiments is carefully crafted to ensure that participants fully understand their tasks and to get a very specific kind of response from them. To encourage participants to formulate their decisions thoughtfully and to do the best they can, they are rewarded (typically paid in cash) according to their performance in the experiment.

The earliest experiments in development economics sought to understand how farmers in India make risky decisions in order to know whether risk aversion among poor farmers might prevent them from trying new seeds. As we'll describe in chapter 12, World Bank economist Hans Binswanger[19] implemented simple risk "games" in which farmers were given money and had to decide how much they were comfortable risking on defined gambles. By presenting all farmers with the same series of gambles and putting real money on the table, economists can measure each farmer's degree of risk aversion, which can influence many real-world decisions and thereby inform the design of development policy and interventions.

Natural Experiments

The international migration of labor is an important component of globalization and economic development in many LDCs. The number of international migrants, or people residing in a country other than their country of birth, has increased at an increasing rate over the past forty years, from an estimated 76 million in 1965 to 215 million in 2010. According to the World Bank, migrants sent US$325 billion home to LDCs in 2010, far more than official development assistance programs did: for each dollar of aid rich countries give poor countries, migrants send home more than $2.50. The flow of international migrant remittances to LDCs is increasing—faster than the number of migrants, in fact.

How do these remittances affect migrant-sending economies? This is an important question in development economics. Unfortunately, the selection problem makes it very hard to answer. Migration is not like a random treatment; households and individuals decide whether or not to migrate. There is a huge selection problem. The households that send family members off as migrants are different in many ways from the households that do not, in ways that are likely to affect almost any outcome we want to study.

We could imagine a hypothetical thought experiment in which we randomly plucked some people out of some households and made them migrate (the "migration-treatment" group), while keeping everyone else at home (the control group). Then we could go back at some future date and compare outcomes of interest, like remittance income, kids' schooling attendance, and productive investments, between the two groups.

Sidebar 2.5 A Remittance "Natural Experiment" from the Philippines

It's hard to imagine designing a randomized control trial to evaluate the impacts of migrant remittances, but Dean Yang came up with what might be the next best thing. He noticed that, at the moment of the Asian financial crisis of 1997, Philippine households had migrants in many different Asian countries. In some cases, the same household had migrants in more than one country. When the crisis hit, the Philippine peso devaluated more against some Asian countries than others. When the peso devalues, the value of remittances in pesos increases. For example, each Hong Kong dollar a migrant sent home turned into more Philippine pesos than before the crisis.

Nobody expected the crisis to happen. The impact on each household's remittances depended on where its migrants happened to be at the time of the crisis. That, Dean argued, makes the changes in remittances almost as good as random.

He found that a 1% peso devaluation increased remittances by 0.6%. These positive remittance shocks caused households to invest more time and money in human capital as well as in local businesses. Child schooling rose, while child labor decreased.

This study was important because of its "natural-experiment" approach to measure remittance impacts and its finding that remittances have a positive impact on investments in migrant-sending households.

Dean Yang, "International Migration, Remittances and Household Investment: Evidence from Philippine Migrants' Exchange Rate Shocks," *Economic Journal* 118 (April 2008):591–630.

With a randomized "migration treatment," households with and without migrants would be the same, on average, except for migration. There would be no selection problem.

Such an experiment, of course, is unrealistic, and even if it weren't, it would be unethical to make some people migrate (even if they didn't want to) while preventing others from migrating (even if they wanted to). It would violate the "do no harm" axiom, which we'll learn about later in this chapter. There is no RCT to study the impacts of migration on migrant-sending economies.

Is evaluating migration's impacts hopeless, then? Dean Yang, an economist at the University of Michigan, found a way (see sidebar 2.5).

Whereas with laboratory experiments researchers have direct control over the experiment and with RCTs researchers typically have indirect control over the experiment (because a partner NGO or agency typically administers the "treatment"), with natural experiments researchers have no control whatsoever. Instead, they take what history, legislation, or nature serves up and try to uncover circumstances in which people had little choice about being "treated" with something of interest. Used in the right way, such circumstances can remedy what is known as a "reflection problem." Often, we want to know how some "treatment" *X* (like remittances, in the example we just saw) affects some outcome *Y* (like poverty), but *Y* may also affect *X*. Poor households might be more likely to migrate in search of higher incomes, or they might be less likely to migrate if migration involves high costs and risks. If *X* reflects *Y* in this way, it becomes very difficult to disentangle the effect of *X* on *Y* from the effect of *Y* on *X*. The selection problem is related to this reflection problem.

Hollywood gives us a nice illustration of the reflection problem. Some famous movie scenes with villains have a hall of mirrors. Every time the villain moves, so does his reflection in a bunch of different mirrors. That's what happened to James Bond in *The Man with the Golden Gun*. This is a classic identification problem. You see the outcome (all those reflections of the bad guy raising his gun), but you don't know the cause (how can you identify the *real* bad guy who makes all the reflections move?).

That's how it often is with identifying cause and effect in economics. We see the outcome, but usually we don't have a neat RCT, so we need more information to figure out the cause. For example, if the villain coughs or steps on a twig, James can isolate him from the reflections and take him down. The sound is associated with the real bad guy but not his reflections, so it lets Bond figure out which is which. Basically, that's the strategy we have to follow in order to establish cause and effect in economics when we don't have a good RCT.

Economists are always on the lookout for variables that are correlated with treatments but not with the outcomes they study. These are called "instrumental variables." An example is the Asian economic crisis in Dean Yang's study (sidebar 2.5). Many of the most important development economics questions cannot be studied with the aid of well-designed RCTs. Econometric methods are then used, along with carefully chosen instruments, in an effort to isolate cause and effect. In the rest of this book we will learn about a number of different studies in

which economists came up with novel ways to identify impacts without the benefit of an RCT.

James Bond found a more straightforward solution to his identification problem. He quickly shot out all the mirrors until the only thing left was the bad guy, Francisco Scaramanga!

THE ETHICS OF EXPERIMENTS

Experimenting on people raises ethical considerations. History gives us extreme and frightening examples of incidents in which people have been harmed by research, particularly in the medical and psychological areas. They include deliberate infection with serious diseases, exposure to biological or chemical weapons, human radiation, and many other atrocities. Some are less obviously harmful. A Stanford University study funded by the US Office of Naval Research in 1971 used students as guinea pigs to investigate the causes of conflict between military guards and prisoners. Students participated voluntarily for $15 per day. They were randomly assigned to play the roles of prisoners and guards in a mock prison in the basement of the psychology building, but they internalized their roles too well. By the time the experiment was terminated, the guards were subjecting their prisoners to physical and psychological abuse. The Stanford Prison Experiment often is held up as an example of unethical scientific research.

Today, any time human subjects are part of research, careful measures are required. Institutional review boards (IRBs) have to approve, monitor, and review biomedical as well as behavioral research involving humans. IRB approval is even required in order to carry out most kinds of economic surveys, because when you ask people questions in a survey, the respondents are your research subjects. It is important to remember this anytime you engage in social science research involving people. Guidance on complying with human subjects requirements is available at most universities and from the US Department of Health and Human Services (HHS; www.hhs.gov/ohrp/archive/irb/irb_guidebook.htm).

Despite IRB reviews, as RCTs have become a dominant methodology in development economics, they have raised considerable controversy, including with regard to ethics. Economists Chris Barrett and Michael Carter point out four classes of ethical considerations that arise in experiments by development economists.[20] These are discussed in the following four subsections.

Adverse Consequences of Experiments

The first rule in studies involving humans is the "do no harm" principle. Experiments manipulate people's environment in an effort to learn about their behavior. If in doing so they harm people, they are unethical and should not be implemented. This is the primary focus of IRBs.

Often, adverse effects of experiments are predictable and clear-cut. For example, if an RCT would encourage people to do something illegal or would put them in harm's way, it is definitely not ethical. An RCT in India created incentives for people to get driver's licenses without necessarily successfully completing the required training and testing. This potentially put innocent people at risk on the roads.

Other experiments are less blatant but still raise concerns. For example, researchers in China studied the impact of treating kids for iron deficiency (anemia) on school performance. Some children known to have anemia were given iron pills, and others were not. This study would not be approved in the United States because withholding treatment for something like anemia would not be considered ethical.

Barrett and Carter listed a number of cases in which experiments are likely to produce adverse consequences. One experiment tested whether large grants of money to women's organizations changes them in ways that lead to the exclusion of poor women, potentially harming poor women. The study's finding that it did lead to exclusion seems to confirm that poor women may have been harmed by the experiment.

Think about the credit experiment we looked at previously, in which some people with low credit scores were given loans. Does it comply with the "do no harm" rule? Fannie Mae (the Federal National Mortgage Association), a US-government-sponsored enterprise, made many home loans to people who should not have gotten them. This was a major cause of the "Great Recession" beginning in 2008. Needless to say, it was not a good thing for the people who shouldn't have gotten loans and ended up going into default. Giving loans to people who do not qualify for them can put their property and reputation at risk.

It is hard to imagine doing any harm by giving people good stuff like goats, chickens, or cash. Yet as we have seen, cash transfer programs can potentially harm some nonparticipants, for example, by pushing up local prices for food and other items they buy. This is not to say that these programs should not be implemented—they almost certainly do considerably more good than harm. Nevertheless, when we implement experiments or other programs, we have a responsibility to anticipate

possible negative impacts on participants or nonparticipants and do whatever we can to mitigate them. This is part of the "do no harm by doing good" maxim.

Informed Consent

There is a difference between people being willful participants in experiments and people as subjects manipulated for research ends. The right of informed consent is well accepted; everyone who participates in a drug trial does so voluntarily. In RCTs, people often are unaware that they are (or are not) part of an experiment. IRBs require that participation in research studies, including simply being surveyed as part of an RCT, be strictly voluntary. The question, then, is how much information researchers should give their human subjects before they decide whether or not to be part of an RCT.

Blindedness

In medical research, people can know they are in an experiment without knowing whether they get treated. The use of a placebo makes this possible: the placebo pill looks the same as the real thing, so no one except the researcher knows who's being treated. Very few RCTs attempt to use a placebo. (For an interesting exception see sidebar 2.6.)

When you give someone an economic treatment, it's hard to keep it a secret. If a person knows she is in an RCT but ends up in the control group, she knows it. Keeping who gets the treatment and who doesn't a secret is a basic tenet of medical research, but it generally is not possible in economic RCTs.

The most important ethical rule, we have seen, is to do no harm. If people know they are in the control group, might they suffer emotional distress because they are not getting the benefits of the treatment? Imagine that you are desperately poor and malnourished. Could there be adverse emotional, psychological, even health consequences of knowing that you have been excluded from a treatment that could significantly improve your situation?

If so, there could be not only ethical but also research concerns. If you know you're in the control group and lose hope as a result, you could end up doing worse than you would have done without the experiment. Thus, the treatment group might look better off compared to you, making it seem like the treatment worked better than it did.

Sidebar 2.6 What? An Economic Placebo?

Using a placebo is basic in medical research. Treatment and control groups take an identical pill, but no one (except the researcher) knows which pill is the real thing. That's important, because if you know you are (or aren't) getting the treatment, the experiment is likely to get contaminated; for example, your behavior might change (like taking an antihistamine and then riding off into the pollen on your bicycle).

Economic RCTs are different, though. For example, people know whether or not they're getting a cash transfer. It's impossible to give people an "economic placebo."

. . . or is it? Four researchers ran a RCT in which farmers didn't know whether or not they were getting the real treatment. In randomly chosen treatment villages, farmers got a modern high-yielding variety (HYV) of cowpea seed. In control villages, they got the placebo: a traditional variety (TV). None of the farmers knew which seed they were planting. The result? Yields were the same between the treatment and control groups.

In another set of treatment villages, the farmers ran a normal RCT: the farmers knew they were getting the HYV. When the farmers knew they were planting the new seed, their yields were significantly higher. When people knew they were getting the real treatment, their behavior changed. That is a placebo effect.

High-yielding seeds are designed to produce a bigger harvest when combined with the right combination of inputs: fertilizer, water, and so forth. The experimenters could have given farmers a package of inputs to use along with the seeds. That way, the only difference between treatment and control farmers would have been the seed, itself. Scientists frequently run experiments in which they control all inputs on experiment station plots. They are left with the question: Do experiment station results reflect what really happens out on farmers' fields? If we want to find out how a new seed affects crop yields in the real world, we have to recognize that farmers' input choices will be a key factor shaping the outcome—and those choices will depend on knowing which seed they're planting.

Erwin Bulte, Lei Pan, Joseph Hella, Gonne Beekman, and Salvatore di Falco, "Pseudo-Placebo Effects in Randomized Controlled Trials for Development: Evidence from a Double-Blind Field Experiment in Tanzania," *American Journal of Agricultural Economics* (in press; http://ajae.oxfordjournals.org/content/early/2014/03/19/ajae.aau015.full).

Is it ethical to involve people in experiments without their knowledge? If you answer "no" to this question, then you immediately hit another one: Can you reliably measure the impact of a social treatment if people *know* whether or not they are in the experiment? Will people—even people in the control group—change their behavior in ways that tarnish the RCT?

Targeting

Development organizations and governments have scarce resources to carry out development projects. It might seem logical (and ethical), then, to efficiently target these resources. Community knowledge can be used to make sure help goes to those most in need. RCTs routinely treat individuals who are not most in need of the treatment, while denying treatment to those who are. Strict randomization thus is viewed by many as being both wasteful and unfair. This can—rightly—be a stumbling block to convincing governments and communities to participate in RCTs.

An additional concern in experiments is the Treat-and-Run Syndrome. PROGRESA left Mexico with one of the world's most comprehensive social welfare programs, one that continues to this day. However, most RCTs are not conducted as part of large-scale government programs; many researchers abandon their research sites once the results of their RCTs are in. What are the long-term impacts of this "treat-and-run" way of doing research? If providing benefits to a treatment group does no harm during an experiment, does ceasing those benefits do no harm in the long run? What are the effects of leaving people behind after an RCT is over?

However you might answer these ethics questions, on one thing we can all agree: anytime we use human beings as research subjects, we have a special responsibility to make sure that we do them no harm, not only during the experiment, but afterward, as well.

THE INVARIANCE ASSUMPTION

Experiments, in order to be valid, must satisfy the *invariance assumption,* which states that the actual program will act like the experimental version of the program. Often, the purpose of RCTs is to test interventions that, if deemed successful, will be scaled up to a larger—or

perhaps the entire—population. Will the large-scale program have the same kinds of impacts as the small-scale RCT? Or is there something about ramping up a project that creates new impacts not captured in experiments?

Actually, there may be. Once the program gets scaled up, the control group disappears. Linkages can transmit impacts of the program through the whole economy. Now everyone is likely to be affected, directly or indirectly, by the treatment. We call the total effect on the economy the "general equilibrium (GE) effect." We look at GE effects of projects at the end of this chapter. GE effects are a major reason why the invariance assumption may be violated. An intervention does not have to be particularly large in order to unleash GE effects; it only has to be important relative to the size of the economy in which it happens. In a poor region, a small project can have a large GE impact.

MULTIPLE TREATMENTS AND INTERRELATED OUTCOMES

In the worms experiment there was a clearly defined treatment (for worms) and outcome of interest (children's school attendance). Often, programs have multiple instruments (e.g., a cash transfer plus conditionality and eligibility requirements, or cash transfers and input subsidies or crop-price supports) and interrelated outcomes. In these cases, it quickly becomes difficult to connect specific components of the program with specific outcomes of interest.

Consider the social cash transfer (SCT) program initiated in 2011 in Tigray, Ethiopia. Many of the households eligible to receive the SCT already participated in a different transfer program: the Productive Safety Net Program (PSNP) had been offering them the opportunity to work a limited number of days on public projects in return for food or cash. When a household gets the SCT treatment, it stops getting the PSNP one. The new program crowds out the old, and both coexist within the same (treatment and control) localities. Disentangling the effects of these two programs is essential if we wish to evaluate the SCT's impacts. Some of the best experiments involve multiple treatments, but when there are many different interventions happening simultaneously, RCTs may not be up to the task of sorting out the impacts.

The impacts of most projects and policies are almost certain to be heterogeneous, with both winners and losers. Few experimental studies consider the ways in which some people may gain while others may lose as a result of a policy or program.

"WHETHER," "WHY," AND "HOW"

Consumer theory gives us a familiar equation relating a household's demand for a good *(D)* with its income *(Y)* and the market prices of this and other goods *(P):*

$$D = \beta_0 + \beta_1 Y + \beta_2 P$$

This is what we call a structural equation. It is structural because it is derived from a theory of how the household economy works. Thanks to consumer theory, we know why income and prices are in this equation, and we even know what signs to expect on the parameters (for example, $\beta_1 > 0$ if we are dealing with a normal good, and $\beta_2 < 0$ if P is the price of the good in question).

When it comes to estimating this equation, though, we have a problem. Current income is endogenous. It is the result of work and other choices people make, and those choices might be related to consumption decisions in ways other than through income. Thus, when we compare demands among people at different income levels, there is likely to be a selection problem.

In chapter 6 we'll learn about Jacob Mincer, who argued that people's permanent income depends on their schooling *(S)* and work experience *(E),* which we can treat as given at any point in time:

$$Y = \alpha_0 + \alpha_1 S + \alpha_2 E + \alpha_3 E^2$$

We could substitute this equation into our consumer demand model, eliminating the problem income variable and expressing demand as a function of schooling, experience, and prices:

$$D = \gamma_0 + \gamma_1 S + \gamma_2 E + \gamma_3 E^2 + \gamma_4 P$$

This is what we call a "reduced-form model." In economics, a reduced-form model is what you get once you've solved for the endogenous variables (here, income). In the reduced-form model, the variable of interest (here, D) is a function only of exogenous variables. If you do the algebra, you'll find that its parameters are functions of the parameters in the other two equations.

We might use econometrics to estimate this reduced-form model with survey data. We might find, for example, that the demand for smartphones increases with people's schooling. However, we would not be able to interpret the economic meaning of this result without

knowing the underlying structural model. There are many reasons why schooling might influence the demand for smartphones. According to the structural model, schooling increases income, which in turn increases cell phone demand. A finding that schooling positively affects cell phone demand would be support for the hypothesis that cell phones are normal goods, based on the structural model.

A common rap against experimental methods is that they are reduced form. In a well-designed experiment, the treatment is exogenous. We estimate its impact on an outcome of interest. Experiments are a good way to test whether a treatment has an effect, but like other reduced-form methods, they do not tell us why. The economist Angus Deaton wrote: "In ideal circumstances, randomized evaluations of projects are useful for obtaining a convincing estimate of the average effect of a program or project. The price for this success is a focus that is too narrow to tell us 'what works' in development, to design policy, or to advance scientific knowledge about development processes."[21]

Designing good policies depends on understanding "why" as well as "whether." It also requires focusing our research on the highest-priority questions.

The best experimental studies not only test program impacts but also try to offer glimpses into the structural reasons why a treatment produces the outcomes it does. For example, in a clever experimental study in Kenya, some farmers were offered free fertilizer delivery early in the season and others not, while still others were offered a fertilizer subsidy. The study found that offering delivery early was more effective at increasing fertilizer use than was a subsidy.[22]

In general, though, it is far more difficult to answer the question of *why* a treatment has the effect it does than *whether or not* there is an effect and *how big* the effect is.

OPPORTUNITY COSTS

So your RCT finds that a program is effective at achieving its goals. Should the program be scaled up? The answer implicit in most experimental studies seems to be "yes." But is it the best way? Economists often talk about "opportunity costs." The opportunity cost of doing one thing is the value of what you could have done instead. When doing RCTs, it is easy to forget that every project and every way of carrying out a project has an opportunity cost. Finding that a treatment has a significant effect on an outcome of interest does not necessarily mean

that the treatment is the best use of scarce public resources. A cash transfer, output price support, technology policy, or fertilizer voucher all might raise incomes in the beneficiary households, but they are unlikely to be equally effective at transforming a dollar of public expenditure into an increase in income in the treatment (or nontreatment) households.

COST-BENEFIT ANALYSIS

Economics offers a methodology to choose among different actions. It is called cost-benefit analysis (CBA, for short). You probably use CBA all the time without even thinking about it, like when you picked up this development economics book! CBA is the basic tool that development banks use to determine whether a development project is viable before it gets funded, and it can be used to compare the viability of different projects, as well—provided that the costs and benefits of projects can be quantified.

The basic idea behind CBA for development projects is simple: add up all the benefits and costs of the development project and take the difference. If this difference is positive, the project is viable; if not, then there is not an economic basis for undertaking the project. If it is positive for two or more different projects but you can only afford to carry out one of them, pick the project in which the difference between the benefits and costs is greatest.

In practice (like everything in life, it seems), CBA gets complicated. For one thing, most projects involve heavy start-up costs in the short run and benefits that are in the future. A dollar in the future is not worth the same as a dollar today—that's why banks have to pay interest in order to get us to save.

Discounting and Net Present Value

CBA has a straightforward way to deal with the timing problem: use the interest rate to discount future values and express them in present value (PV). If i is the interest rate, the PV of \$100 of income a year from now is $\$100/(1 + i)$. If the interest rate is 5% (that is, .05), we get $\$100/1.05 = \$95.24$. If you had \$95.24 today, you could turn it into \$100 a year from now by putting it in the bank at 5% interest—and waiting.

What is the PV of \$100 two years from now? In other words, how much would you need to put in the bank today to end up with \$100 after two years? The answer is $\$100/(1 + i)^2$. At a 5% interest rate, the PV of \$100 two years from now is $\$100/(1.05)^2 = \$90.70$. When doing CBA,

we convert all future benefits and costs to PV by dividing them by $(1 + i)^t$, where t is the time period: $t = 1$ in year 1, $t = 2$ in year 2, and so on.

Once we have discounted all future benefits and costs of a project, we sum their differences to get the project's Net Present Value (NPV):

$$NPV = \sum_{t=0}^{T} \left(\frac{Benefits^t - Costs^t}{(1 + i)^t} \right)$$

The capital Greek sigma (Σ) denotes the sum; our formula adds up the discounted difference between benefits and costs from the start of the project ($t = 0$) until the end of the time period over which we wish to perform the cost-benefit analysis ($t = T$).[23]

The NPV formula is the basis for carrying out any CBA. If NPV > 0, the project passes the economic cost-benefit test. If you can only fund one project, on purely economic grounds choose the one with the highest NPV.

Determining Benefits and Costs

The trick always is in figuring out what the benefits and costs are. Often, a project's costs are immediate and known. For example, it is not hard to determine the cost of running an extension program to train one hundred farmers on how to use a new technology and giving each farmer a technology start-up package (for example, high-yielding seed and fertilizer to plant one acre). Or the cost of building a new school room and staffing it with a teacher. Or of carrying out an immunization program in one hundred villages.

Calculating benefits can be a different matter, though. If you carry out the extension program, how much higher will the farmers' incomes be? If you build the school, will students attend? How many parents will bring their kids to the clinic to get the immunization? If more kids attend school or get immunized, will their future incomes go up because they become more productive?

This is where experiments and the other evaluation methods in this chapter can help. Cleverly designed RCTs can provide estimates of how many farmers will adopt a new technology and how much their yields are likely to increase if they do. The PROGRESA and African cash transfer RCT studies described earlier in this chapter estimate impacts on school attendance. Chapter 6 includes an RCT to evaluate the demand for immunizations.

TABLE 2.1 PRESENT VALUE OF COSTS AND BENEFITS OF A HYPOTHETICAL PROJECT

t	Cost*(t)*	Benefit*(t)*	Benefit*(t)*–Cost*(t)*
1	100		–100
2		18.18	18.18
3		16.53	16.53
4		15.03	15.03
5		13.66	13.66
6		12.42	12.42
7		11.29	11.29
8		10.26	10.26
9		9.33	9.33
10		8.48	8.48
NPV: 15.18			

Here's a simple illustration of the mechanics of CBA: imagine a project that would cost $100 to carry out, with all of those costs occurring in year 1. Beginning in year 2, based on our experimental or other estimates, we expect the project to produce benefits of $20 per year. Your funding agency requires that the project break even within ten years—that is, the project's NPV, evaluated over a ten-year period, must be positive. Is it?

Using a fairly conservative (10%) discount rate, we can construct a table showing the PV of this project's costs and projected benefits (table 2.1).

The balance of annual benefits and costs starts out negative, because there are only costs in year 1. It turns positive once the project begins to yield the $20 benefit per year. The $20 number doesn't appear in this table, though. Benefits have to be discounted: the PV of $20 after one year is $18.18; after two years it is $16.53, and so on. (If there were costs in years 2–10, they would have to be discounted, too.)

Adding up all of the discounted benefits and costs over the ten-year period, we get an NPV of $15.18. It is greater than zero; thus, the project is economically viable. Whether it is economically optimal will require comparing this to the NPVs of competing projects.

You might want to experiment using the CBA worksheet posted online for this chapter. You would find that this project would not pass the economic cost-benefit test if the benefits were $17 per year, if the interest rate were 14%, or if there were an annual cost of $2 to keep the project going.

In chapter 6 we carry out a simple cost-benefit analysis of going to school for a child in a poor Lesotho village. It is not unlike the CBA you might have carried out while deciding whether or not to study development economics.

Non-Economic Benefits and Costs

Many benefits and costs cannot be quantified. CBA can be a good tool for evaluating economic costs and benefits and selecting projects on economic grounds. Clearly, there are reasons to carry out projects on other grounds, as well. How can one deny a child education or good health if it is at all possible to provide her with these basic human rights? Non-economic benefits strengthen the argument for carrying out some development projects. Non-economic costs can do the opposite. An example of the latter is an activity that produces negative externalities, for example, a negative environmental impact. Positive externalities, on the other hand, can strengthen the case for a project. For example, "treating" some farmers with information about better cultivation practices could have positive externalities if the "treated" farmers share this information with others. In short, CBA is a useful economic tool, but it may not be the sole criterion for implementing a development project.

BEYOND EXPERIMENTS: LOCAL ECONOMY-WIDE IMPACTS OF DEVELOPMENT PROGRAMS

Suppose we wish to evaluate the impact of an income transfer program on rural poverty. Poor households receive the transfer, which might entail some sort of conditionality (for example, PROGRESA's requirement that children attend school) or not (the case in almost all of the SCT programs in Africa). Figure 2.1 illustrates the pathways by which this project might impact a local economy. Arrow (a) represents the transfer's direct effect on the income of a recipient (poor) household. This is equal to the amount of the transfer. With higher income, the household's demand for normal goods and services increases. The transfer can affect the household's production activities in a number of different ways. By raising the household's income, it can stimulate consumption demand, including the demand for leisure and goods produced by the household.[24]

For example, an increased demand for food could encourage a subsistence household to grow more food crops, while an increased demand

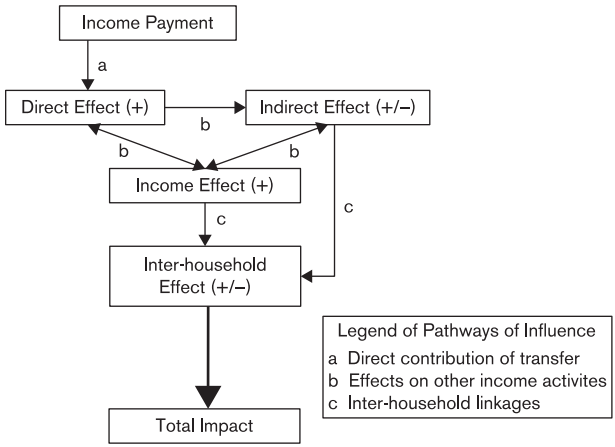

FIGURE 2.1. An income transfer project creates both direct and indirect income effects in the treated economy.

for leisure could do the opposite. If leisure demand increases, the household's wage income could fall. The transfer could loosen liquidity constraints on crop production, enabling the household to purchase more fertilizer and other inputs or shift into input-intensive cash crops.[25] Finally, it could reduce income risk, and this might encourage the household to invest more of its scanty resources in risky activities. (We'll learn about agricultural household behavior in chapter 9.) Conditionality could create still other impacts in recipient households. For example, the requirement that children attend school could decrease the family's labor available for crop production.

Arrow (b) depicts these myriad indirect effects of the transfer on the treated household's income from production and labor activities. Experimental methods, when feasible and carefully executed, can provide insights into the net influences represented by arrows (a) and (b).

As the recipient households demand more consumption goods and change their allocation of resources to production and wage activities, others in the local economy invariably are affected. Local markets transmit impacts of the transfer from the recipient to nonrecipient households, as represented by arrow (c) in the figure. Households and businesses supplying goods and services to the recipient households benefit. If the transfer alters the recipient household's wage labor supply, this could drive up wages, or as consumption demand rises, so might local prices. These will affect nonrecipient households in other

(possibly negative) ways. As local activities adjust, a new round of changes in input demands, incomes, and household expenditures follows, creating additional rounds of changes in incomes and expenditures. Given income leakages, successive rounds of impacts become smaller and smaller, and the total (direct plus indirect) effect of the program eventually converges to an income multiplier. To the extent the goods demanded by the recipient households are supplied locally, the income transfer could create a multiplier considerably greater than one. On the other hand, if the recipient households purchase goods from outside the local economy, some or perhaps most of the multiplier will go elsewhere. Clearly, the behavior of the households that get the transfer to begin with is critical in shaping the impacts that result from the program, but so is the behavior of the nontreated groups and the structure of the local markets connecting them with each other.

The economic linkages that transmit impacts through economies are called "general-equilibrium feedback effects." In a few cases, RCTs have collected data on ineligible households and found evidence that they are affected by treatments. One such study was done on the effects of Mexico's PROGRESA on the households that did not get PROGRESA transfers. The impact was found to be positive, implying that only focusing on the treated underestimates the program's impact.[26]

To understand the ways in which a treatment affects both treated and nontreated households, we generally have to go beyond RCTs and try to model economic linkages. As the diagram in figure 2.1 illustrates, the direct and indirect impacts of an intervention are shaped by how households change their supply and demand decisions and by the structure of local markets, which in turn reflect various constraints (technology, transaction costs, liquidity, risk). Performing project evaluations in such environments may require integrating models of heterogeneous households into a model of the whole local economy, a local general-equilibrium (GE) model. A model for the economy targeted by the project (village, region, rural sector) can provide a laboratory in which the project is designed and its impacts assessed, using a simulation approach.

There are fundamental differences—some might call these philosophical differences—between RCTs and simulation models. An RCT, we've seen, is like a drug experiment; it can tell us whether something works, but not why. An advantage of experiments is that statistical significance can be attached to RCT findings; for example, "with 95% certainty we can say that the transfer increased food consumption by

between $10 and $15 per month." The validity of an RCT depends on getting the experiment right; otherwise, the findings, however significant statistically, may be biased. At conferences where researchers present studies using RCTs, much of the discussion centers around what might have gone wrong in the experiment and how this might have affected the results. This illustrates how much more difficult it can be to run "clean" economic experiments than drug trials. Often, one is left with questions about why a treatment had the effect that it did on those who got the treatment.

Simulation models try to answer the question "why" while capturing complex interactions that shape project outcomes, in ways that often are beyond the reach of experiments. The validity of a simulation model depends on getting the model right. Imagine a flight simulator. Schools do not teach pilots how to fly by hitting them with dangerous real-world situations in mid-air. Pilots can step into a flight simulator. The simulator is programmed with equations representing the physics of flight. It becomes a laboratory in which flight experiments are conducted. If you've ever played a computer game, you know what simulations are all about. If the flight simulator is programmed wrong, well, you won't want to fly with that pilot!

A simulation approach to project impact evaluation highlights the interactions within the local economy that transmit impacts, good or bad, from directly affected actors to others in the economy. We can construct simulation models using data from the same surveys that are used to do RCT research. If our simulation model represents the way in which the local economy works, it can be a valuable tool to understand the full, economy-wide impacts of cash transfers and many other programs and policies.

There are two main knocks against simulation models. One is that they depend on getting the model right, especially how agents behave and how markets transmit impacts from one agent to another (like getting the flight simulator equations right). Another is that it is more difficult (though not impossible) to attach statistical significance to simulation results.

A new method, local economy-wide impact evaluation (LEWIE), uses data from RCT surveys to estimate simulation models and construct confidence bounds around their results. This is a step in the direction of bringing together the best of RCT and simulation methods.

A LEWIE simulation model was used to evaluate the local GE effects of a cash transfer program in the southern African country of Lesotho.

Sidebar 2.7 Impacts of a Treatment on the Nontreated in Lesotho

When poor people get cash transfers, they spend them. This transmits impacts of cash transfer programs from treated to nontreated households. Lesotho's Child Grants Program (CGP) seeks to improve the living conditions, nutrition, health, and schooling of orphans and vulnerable children. It seeks to accomplish this via an unconditional cash transfer targeted to poor and vulnerable households.

A local economy-wide impact evaluation (LEWIE) found that each $1 transferred to a poor household raises total village income by $2.23, with a 90% confidence interval (CI) of $2.08 to $2.44. Even though all of the cash transfers go to poor eligible households, nearly half of the benefits they create ($1.18) go to ineligible households.

If there are constraints that limit the local supply response, though, higher local demand may push up prices instead of stimulating production. Price inflation reduces the multiplier in real (price-adjusted) terms. (We'll learn how to adjust income for inflation in chapter 3.) It raises consumption costs for everyone in the local economy. If supply constraints are severe, the *real* income multiplier may be as low as $1.36 (CI: $1.25–$1.45). The study found that loosening capital constraints, say, through effective microcredit programs that enable households to buy more crop inputs, is a key to avoiding inflation and raising the real transfer multiplier.

This study is important because it reveals potential impacts of cash transfer programs that are unlikely to be picked up by RCTs—including impacts on households that do not get the cash. Large local income multipliers suggest that social cash transfer programs promote income growth in poor villages. That's good news for both social welfare ministers and finance ministers in LDCs.

J. Edward Taylor, Mateusz Filipski, Karen Thome, and Benjamin Davis, "Spillover Effects of Social Cash Transfers: Lesotho's Child Grants Program," in *Beyond Experiments: Evaluating Development Impacts with Local Economy-Wide Models,* edited by J. Edward Taylor and Mateusz Filipski (Oxford: Oxford University Press, 2014), 181–202.

It uncovered import spillover effects, including effects on the households that did not get the transfer (see sidebar 2.7).

www.rebeltext.org/development/qr2.html
Learn more about what works and what doesn't by exploring multimedia resources while you read.

APPENDIX

The Math of Selection

The math behind RCTs is not very hard, but it takes most people some time to wrap their minds around it because it involves some "what ifs." Here's how it works:

We want to know whether a treatment (like a development project) affects some outcome of interest (say, income or health). Let's call person i's outcome Y_i. If a person gets treated, the outcome is Y_{1i}, and if she does not get treated, it is Y_{0i}. Each person has both a Y_{1i} and a Y_{0i}, but there's a catch: we can only see one of them. If i gets treated, we see Y_{1i} but not Y_{0i}. If she doesn't get treated, we see Y_{0i} but not Y_{1i}.

Let's make a variable D_i that equals 1 if person i gets treated and 0 otherwise. A concise way to represent the outcomes is:

$$Y_i = \begin{cases} Y_{1i} \text{ if } D_i = 1 \\ Y_{0i} \text{ if } D_i = 0 \end{cases}$$

The outcome we "see" for person i is whatever it would be without the treatment, Y_{0i}, plus whatever effect the treatment has, which is $(Y_{1i} - Y_{0i})D_i$. For short, let's call the actual effect of the treatment ρ:

$$Y_i = Y_{0i} + \underbrace{(Y_{1i} - Y_{0i})}_{\rho} D_i$$

The treatment effect, ρ, is what we want to find out. It is the change in the outcome that is *caused* by the treatment. (If the person does not get treated, $D_i = 0$, so this second term is zero.)

Now suppose we simply compare expected or average outcomes for people who do and do not get the treatment. In stats talk, the expected or average outcome given that a person gets the treatment is $E[Y_i|D_i = 1]$, and the expected outcome for people who don't get the treatment is $E[Y_i|D_i = 0]$. ("E" means "the expected value of," and the slash marks mean "given that.") The average difference we see between the people who are treated and the people who are not, then, is

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

This difference is not the average effect of the treatment on the treated, because it includes selection bias. The average effect of the treatment on the treated is the difference between (1) the expected outcome for people with the treatment, given that they got it ($E[Y_{1i}|D_i = 1]$), and (2) the expected outcome for these same treated people *if they had not been treated* (which we can call $E[Y_{0i}|D_i = 1]$). In other words, the average treatment effect on the treated, which is what we want to know, is

$$E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0]$$

Imagine the people who get the treatment (that's the first term). If, after they get the treatment, we could put them into an Orwellian time machine, send them back in time, and then not treat them, we'd have the second term. If that person did not change in any other way, the difference would be the true average effect of the treatment on the treated.

Obviously, we cannot both treat and not treat the same people. We have to compare people who get treated to people who don't. This leaves us with selection bias. Selection bias is the difference between (1) the expected outcome for those who got treated, if they hadn't gotten treated (same as the second term in the expression above: $E[Y_{0i}|D_i = 1]$), and (2) the expected outcome without the treatment for the people who didn't get treated ($E[Y_{0i}|D_i = 0]$). In other words:

$$E[Y_{0i}|D_i = 1]\ E[Y_{0i}|D_i = 0]$$

In the hospital example, the selection bias is negative, because the people who get the "hospital treatment" (the first term above) are less healthy, on average, than the people who don't get the treatment (the second term). It is reasonable to expect that, on average, the people who went to hospital would have had poorer health without going ($E[Y_{0i}|D_i = 1]$) than the people who didn't go got by not going ($E[Y_{0i}|D_i = 0]$). The people in this last group probably didn't go because they didn't need to.

To sum it all up:

$$\underbrace{E[Y_i|D_i = 1] - E[Y_i|D_i = 1]}_{\textit{Observed difference in average health}}$$
$$= \underbrace{E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]}_{\textit{Average treatment effect on treated (positive–we hope!)}}$$
$$+ \underbrace{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{\textit{Selection bias}}$$

What makes this challenging to understand is that the repeated term on the right-hand side of the equation above, $E[Y_{0i}|D_i = 1]$, is hypothetical. We cannot see what did not happen.

Randomization solves the day. If the treatment is truly random, then on average the people who get it are identical to those who do not, so their outcomes without the treatment, on average, are the same: $E[Y_{0i}|D_i = 1] = E[Y_{0i}|D_i = 0]$. The selection bias term disappears, leaving only the average treatment effect on the treated. That's why, in a well-designed RCT, we can estimate the average effect of the treatment on the treated simply by comparing average outcomes for the random treatment and control groups.