

---

# 1 The Experimental Approach



This chapter provides an example of how a randomized evaluation can lead to large-scale change and provides a road map for an evaluation and for the rest of the book. The modules in this chapter are as follows:

## MODULE 1.1: The Power of Randomized Evaluations

### MODULE 1.2: A Randomized Evaluation from Start to Finish

---

## MODULE 1.1 The Power of Randomized Evaluations

*This module provides an example of how a small nongovernmental organization, by subjecting its program to rigorous evaluation, can generate evidence that can change the lives of millions of people.*

In 1994 I went with Michael Kremer to visit the family he had lived with for a year in rural Kenya.<sup>1</sup> We met up with many of Michael's old friends, including Paul Lipeyah, who told us of the work he was doing with International Child Support (ICS) Africa, a nongovernmental organization (NGO) helping government schools in Busia, a neighboring district in Kenya's Western Province. Paul asked us what advice we might offer for improving the effectiveness of ICS programs. Could Michael help evaluate what they were doing? Michael suggested randomized evaluation: if ICS wanted to understand the impact of their programs, they could randomly choose the schools in which they worked and the order in which they phased in new programs.

1. The first-person reflections in this chapter are those of Rachel Glennerster.

Over the following years, ICS randomly evaluated many approaches to improving education outcomes, including providing additional inputs (classrooms, textbooks, teachers); reducing the cost of going to school (providing free school uniforms and school meals); and offering performance incentives (merit scholarships for girls, bonuses for teachers who attended school regularly). Sometimes the programs had the expected impact, and sometimes they did not. But ICS, in partnership with a growing number of researchers, kept innovating and testing in the areas of education, agriculture, women's empowerment, clean water, and health. Their knowledge of how to improve lives and how to evaluate programs kept growing.<sup>2</sup>

In 1997, ICS (in line with World Health Organization recommendations) phased in a program of treating children en masse for intestinal worms (such as hookworm and schistosomiasis). The results were astonishing. Deworming reduced the absenteeism of children in local schools by 25 percent, making it the most cost-effective program for increasing schooling that ICS had tried. Long-term results showed that women who had been dewormed as girls received more education and were more likely to grow cash crops, whereas men who had been dewormed worked 3.5 hours longer per week and were more likely to hold manufacturing jobs and earn higher wages.<sup>3</sup>

On the strength of the evidence, in 2009 Prime Minister Raila Odinga announced a program to deworm 3 million of Kenya's most at-risk children. In 2012 the program was expanded to include pre-school children, in part on the basis of further evidence of cognitive gains for young children from deworming.<sup>4</sup> In 2013, programs to deworm 40 million children are being implemented in Kenya and around the world. ICS could never have hoped to reach so many children with their own programs, and yet through the influence of their evaluation they have helped millions of children in Kenya—and around the world.

Since 1994 we have learned a lot about which programs work and which do not and also about how to run randomized evaluations.

2. The evaluation team at ICS eventually split off from ICS and became Innovations for Poverty Action, Kenya.

3. Studies of this program by Sarah Baird, Joan Hamory Hicks, Michael Kremer, and Edward Miguel are summarized as Evaluation I in the appendix.

4. Owen Ozier, "Exploiting Externalities to Estimate the Long-Term Effects of Early Childhood Deworming," working paper, University of California, Berkeley, 2001.

Until that point, most randomized evaluations of social programs were performed in rich countries in partnership with governments and at very high cost. But the experience of partnerships between researchers and organizations such as ICS showed that it was possible to conduct high-quality randomized evaluations with small organizations on limited budgets and in very poor settings. Although many government programs are being evaluated with randomized evaluations, NGOs have also proved to be flexible and innovative partners in this effort to learn. The challenge of working with new partners on new questions and on modest budgets has spurred innovation in randomized evaluation methodology.

We have learned how to introduce randomization into programs in creative ways, account for and measure spillovers, reliably measure difficult-to-measure outcomes like corruption and empowerment, get the maximum statistical power from a very limited budget, minimize attrition, and design evaluations that answer fundamental questions about why humans behave the way they do and how to motivate changes in behavior.

In this book we have gathered many of the practical innovations from this large body of work. Our goal is to enable more people and organizations to undertake and commission high-quality randomized impact evaluations and thus to build a wider and deeper evidence base on how best to combat poverty. Our hope is that we will see the generation and application of rigorous evidence grow even faster than in the past two decades. By innovating and testing and by feeding the evidence back into even more innovation, evaluators and practitioners can improve the effectiveness of policies and programs and make a real difference in people's lives.

---

## MODULE 1.2 A Randomized Evaluation from Start to Finish

*In this module we provide an overview of the steps in planning and running an evaluation and indicate where these topics are covered in the rest of this book. We tell the story of a randomized evaluation of an education program in India that I experienced firsthand from inception through implementation and analysis to scale-up.<sup>5</sup>*

5. Abhijit Banerjee, Rukmini Banerji, Esther Duflo, Rachel Glennerster, and Stuti Khemani, "Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India," *American Economic Journal: Economic Policy* 2 (2010): 1–30.

### **Starting right: Choosing the right question to test**

Ten years after my trip to Kenya described in Module 1.1, I was working with a group of researchers at the Massachusetts Institute of Technology (MIT) planning our next evaluation. We had recently started what is now the Abdul Latif Jameel Poverty Action Lab (J-PAL) with the objective of promoting the use of randomized evaluations and helping ensure that the results are used to influence policy. Although randomized evaluations (and most of the lessons in this book) are valuable across many disciplines and many regions of the world, our main expertise was in development economics. In prioritizing our work, therefore, we wanted to start by understanding the areas in which rigorous evaluation could be most valuable in informing the debate about poverty in developing countries. Which innovations showed promise but were untested? Which programs were popular with governments and NGOs but had little rigorous evidence to support them?

Community accountability programs were a priority that repeatedly emerged in the conversations we had with organizations working in developing countries as well as in our review of the literature. The enthusiasm for this approach was well articulated in *World Development Report 2004: Making Services Work for Poor People*.<sup>6</sup> The report documented the low quality of services and the lack of accountability, including the chronically high absenteeism of service providers. It argued that community accountability is one of the best ways to improve failing services. The poor who suffered the brunt of the failures were not just more motivated to get services working than were bureaucrats; they were also better positioned because they were right there, at the point of delivery, to monitor the providers. If they were empowered to apply their motivation and monitoring advantages, they would hold providers accountable and services would improve.

In practice, this empowerment took the form of establishing community oversight bodies for schools and clinics and providing communities with information about their rights and the services they should expect. International agencies, NGOs, and governments were all looking to integrate the approach into their work.

There were reasons to think that community accountability would work in practice. Advocates pointed to correlations between active

6. World Bank, *World Development Report 2004: Making Services Work for Poor People* (Washington, DC: World Bank, 2003).

participation of citizens in school and clinic oversight and high-quality services, and cases in which increases in participation were associated with improved services. A popular example was documented in a study from Uganda, in which the government had started disbursing grants directly to schools. A survey found that only 25 percent of these grants were reported as reaching the schools. In response, the government started informing communities of how much money had been allocated to each school, and a few years later, 82 percent of the grants were reported to be reaching the schools.<sup>7</sup>

It was unclear, however, whether the correlation between community involvement and high-quality service outcomes meant that community involvement caused these high-quality outcomes. Communities with high levels of citizen involvement tend to be different from those with low levels of involvement in a number of ways. For example, the town of Brookline, Massachusetts, where I live, has very good public schools and an unusual form of governance in which

---

**CHAPTER 2** explains why it is hard to distinguish the impact of a program from other factors. We discuss alternative approaches the evaluator can use for estimating impact and show how randomized evaluations can help him or her to isolate the causal impact of the program.

---

citizens take responsibilities typically given to full-time town employees. But Brookline also has an unusual concentration of people with high levels of education. People move from miles away and pay high taxes so that their children can attend Brookline's public schools. It is hard to know to what extent the school outcomes are due to the citizen oversight and to what extent they are due to the emphasis on education among local families. More important, it is not clear whether another town encouraged (or mandated) to take up the Brookline model of citizen involvement would achieve the same outcomes.

What about the Uganda example, in which services improved when the information given to local people was increased? Even there it was unclear how big a role empowering communities with information played in the observed changes. Some people believed that the accuracy with which money transfers to schools were recorded had been low during the first year and improved over time. In addition, information on how few of the grant funds made it to the schools was also reported

7. Ritva Reinikka and Jakob Svensson, "The Power of Information in Public Services: Evidence from Education in Uganda," *Journal of Public Economics* 95 (2011): 956–966.

to the Ministry of Education and to donors, and that caused quite a stir. Was it providing information to the ministry and the donors or providing it to the community that caused the change?<sup>8</sup> It's hard to tell.

Because the approach of empowering communities to hold service providers to account was popular and there was little rigorous evidence of its impact, we decided to prioritize this as one of the questions we wanted to test using a randomized evaluation.

### ***Finding a specific context in which to test the question***

Among those keen to both develop and test a community accountability program was Pratham. Pratham is the largest organization, apart from the government, working on education in India. Was India the right context? Was Pratham the right partner?

The education sector in India was plagued by poor public services. The absence rate among primary school teachers was 25 percent,<sup>9</sup> and only 45 percent of teachers present were in the classroom teaching.<sup>10</sup> Pratham had found in other work that even many children who attended school regularly could not read or do simple math by grades 3 and 4.<sup>11</sup> Services were bad and highly centralized; there was scope, therefore, for community accountability to make a difference. The researchers decided that India was a relevant context in which to test a community accountability program.

Pratham was, in fact, an ideal evaluation partner. They knew a lot about education in India, and they wanted to test a community mobilization program. Pratham had previously worked with another J-PAL team, including Abhijit Banerjee and Esther Duflo, to evaluate a remedial education program that trained local young women and placed

8. There is evidence that those schools in closer proximity to a newspaper outlet saw larger improvements in recorded flows of funds, which the authors attribute to greater access to information about the mismatch of funds. However, having a principal and parents with greater access to a newspaper is likely to be correlated with other factors that might lead to greater improvement in recorded flows.

9. World Bank, *World Development Report 2004: Making Services Work for Poor People* (Washington, DC: World Bank, 2003); Nazmul Chaudhury, Jeffrey Hammer, Michael Kremer, Karthik Muralidharan, and F. Halsey Rogers, "Missing in Action: Teacher and Health Worker Absence in Developing Countries," *Journal of Economic Perspectives*, 20 (2006): 91–116.

10. Michael Kremer, Nazmul Chaudhury, F. Halsey Rogers, Karthik Muralidharan, and Jeffrey Hammer, "Teacher Absence in India: A Snapshot," *Journal of the European Economic Association* 3 (2005): 658–667.

11. This study by Abhijit Banerjee, Shawn Cole, Esther Duflo, and Leigh Linden is summarized as Evaluation 2 in the appendix.

them in local schools as tutors for children who had fallen behind. The evaluation had found the program highly effective in helping these children catch up in reading and arithmetic. Pratham wanted to adapt this model to a rural setting and reduce costs by relying on volunteers rather than paid tutors. Successful as their program had been, Pratham also believed that improving the education of children in India required improving the quality of education in the government schools on which most children relied. They had a vision of citizens coming together to ensure that India's children had a better future both by exerting pressure to improve public education and by taking direct action to improve learning. Further, Pratham understood why randomized evaluations were useful, they were in the early design phase of a new program allowing the joint development of the research and project design, and they had the ability to bring a successful program to scale. The research team also recruited Stuti Khemani at the World Bank to bring to the project the expertise and perspective of the World Bank on community accountability programs.

Together, Pratham and the research team decided that Uttar Pradesh (UP) would be a good location for the program and its evaluation. UP is one of India's largest states, with 20 million primary school-aged children. But school quality there was very low: the survey we conducted at the start of the evaluation (our baseline) showed that only 42 percent of 7- to 14-year-olds could read and understand a simple story. Because it would not be feasible to introduce community oversight boards with legal backing to some (randomly selected) schools and not others, we needed a context in which such an oversight board existed but was not very active. UP had legislation mandating that VECs oversee all public schools in a village, but most did not function. Laws existed that gave communities a number of paths to influence the quality of their schools: communities could complain to their members of parliament, local village councils could request funding to hire local assistant teachers, and the village councils had discretionary funds they could use to improve local services. Within UP we chose to pilot and evaluate the program in Jaunpur District, which was close to the state average in terms of literacy and was one of the districts where Pratham was not already working at the time.<sup>12</sup>

12. Some of our qualitative work was done in Gauriganj (Amethi) in the constituency of MP Rahul Gandhi, the son of former Indian prime minister Rajiv Gandhi. However, we became concerned that attempts to put pressure on the education system

---

**CHAPTER 3** discusses how to prioritize questions for an impact evaluation, when nonrandomized methods are sufficient, why understanding the local context is critical in designing an evaluation, and how to choose the right location and partner for undertaking an evaluation.

---

Given our chosen context, we refined our general community accountability question to this: is there a way to mobilize communities to effectively use existing accountability systems in UP to improve educational quality?

***The groundwork: How did we arrive at the final three interventions?***

Over the period of a year, the researchers and Pratham worked together to design the program and the evaluation. Their objective was a program that represented best practices in the area of community mobilization to enhance service accountability and was tailored to the local environment but was replicable. We also had to select which of several alternative versions of the program to rigorously test against each other in the study. Finally, we had to determine how to measure the impact of the program.

**Honing the design of the program**

Both Pratham and the researchers wanted to design and test a program that would be replicable at large scale. In other words, it needed to be relatively inexpensive and not rely on highly trained or educated staff. Scalability concerns also limited the resources Pratham would put into any single village mobilization effort.

We needed to check whether the theory behind the intervention made sense in Jaunpur District. Was the level of learning low? Was there room to improve education (for example, was the rate of teacher absenteeism high)? Did communities have mechanisms they could use to improve education quality? What were the roles and responsibilities of the VECs? If a village wanted an additional assistant teacher or wanted to fire an existing one, how exactly could they do that? How much money did the village council have that could be directed to education? Was there a gap in knowledge that a potential intervention could fill? Did the community know how poor learning levels were? Did everyone know the roles and responsibilities of the VEC? Was

---

to reform might create more response in such a high-profile district than would be typical, and so we moved to Jaunpur.



there relevant information that the village head knew but was not sharing widely (in which case pressing the village head to share that information might be a good strategy)? We spent time in government offices in the state capital of Lucknow finding out exactly what laws were on the books and talked to village heads, teachers, VEC members, students, community members, and national and state education activists.

We found that learning was very poor in Jaunpur, but most people overestimated the level of learning in their community and were shocked when they realized how little their children knew. According to the law, communities had a number of ways they could press for change, most of which ran through the VECs; yet there was very little knowledge of the VECs, what their powers were, how much money they had, or even who was on them. Although village heads tended to know about VECs and could produce a list of the names of the committee members, the committees were usually inactive. In several cases even those whose names were on the lists of committee members did not know of the existence of the committees, let alone that they were on them.

Both Pratham and the researchers wanted to incorporate best practices from other states and countries. The results would be most useful if the evaluation tested a program that the rest of the world would consider a good example of a program designed to mobilize communities to demand quality services. To that end, we studied the *World Bank Participation Sourcebook* as a guide to community participation programs and sought to include all aspects of the guidelines into the program.<sup>13</sup>

The program also needed to work on the ground. So the team spent months going village to village figuring out how best to convey information about both the poor quality of education in the communities and the mechanisms communities had to press for change. How best could one stir up interest in pressing for change? Was it best to have many small discussions or one large meeting that the teacher and the village head attended? How could Pratham best steer the conversation on education away from the perennial topic of the food rations for children attending school and back to the subject of whether children were learning and how to improve that learning?

13. Bhuvan Bhatnagar, James Kearns, and Debra Sequeira, *The World Bank Participation Sourcebook* (Washington, DC: World Bank, 1996).

## Choosing which iterations of the program to test

The researchers and Pratham had many questions about the relative merits of different ways to design a program for community mobilization for accountability, but we knew we had the money and statistical power to compare only a few alternatives against each other. The qualitative work helped us choose which alternatives to focus on.

For example, we had planned to test a very inexpensive intervention that simply provided information on learning levels in the community, the resources available for education, and the mechanisms for generating change on posters distributed throughout the village. But when we put up posters in a community and showed up the next day, we found that most of the posters were gone. This experience did not prove that posters would not work: we had tried using them in just one village. But based on this experience, we thought the chances that this strategy would work were low enough that it was not worth rigorously testing a poster-based intervention. We opted for testing a more interactive approach relying heavily on community meetings to promote information sharing.

One hypothesis to emerge from our qualitative work was that the more actively involved communities were in discovering the inadequate learning levels in their community, the more likely they were to take action to remedy them. Two alternative versions of the program that were ultimately evaluated were designed to test this hypothesis. Pratham developed a testing tool simple enough to be used by community members to determine how many of their children could recognize letters or words and read simple paragraphs or stories. In one arm of the study, communities themselves would generate the information on children's learning and present it at a community meeting at which attendees would go on to discuss what action to take.

We also decided to use the opportunity to test the new version of Pratham's remedial education program (Read India) designed for rural settings. Pratham saw the community mobilization program and the Read India program as natural complements: community mobilization around education and an understanding of levels of reading in the community would be necessary to recruit volunteers to teach children in remedial reading camps after school. From a research perspective, testing community mobilization with and without Read India would help us unpack reasons for the success or failure of community

mobilization. For example, if people did not take action through the public system but did take the opportunity for direct action provided by Read India, it would suggest that the problem was not lack of interest in education or a lack of willingness to take action but rather a lack of faith in the responsiveness of the public system. Similarly, one concern people have about efforts outside the public system is that they undermine motivation to take action within the public system. We would be able to see if action to reform the public system was stronger, weaker, or the same when an option to improve education outside the public system was offered.

### ***The interventions tested***

All three interventions Pratham finally implemented adopted the same basic structure to share information on education and on the resources available to villagers to improve the quality of education.

#### **1. Providing information on improving education services with VECs**

Pratham spent two days initiating small discussions about education throughout the community, culminating in a large communitywide meeting at which teachers and the village head were prompted to provide information about the resources available to improve education in the village, the composition of the VEC, and what resources it receives. Pratham facilitators provided fact sheets and filled in any gaps in information. They also met with every member of the VEC to inform them of their roles and responsibilities.

#### **2. Creating village-based scorecards on reading skills**

The second intervention built on the first, adding the use of a simple tool that Pratham staff taught community members to use to assess the reading outcomes of their own children and the village as a whole. Community members used the tool to generate a village “reading report card” that was then shared with other community members at the village meeting.

#### **3. Demonstrating volunteer-run Read India afterschool camps**

The third intervention supplemented the first and second interventions by providing a way for a motivated citizen to directly improve education levels. Pratham asked for local volunteers to hold afterschool reading camps and trained them over the course of four days in a simple pedagogical technique for teaching reading. The volun-

teers held after-school camps for children who wanted to attend for two to three months, with an average of seven support visits from Pratham staff during that time.

### ***Piloting the interventions***

In addition to the qualitative work Pratham and our research team did to develop the interventions, we also conducted a formal pilot of the interventions in several villages. The pilot had several purposes: it was a final check to see whether the program as designed was feasible and had a reasonable chance of success; it helped us understand in greater detail some of the pathways by which the program might change education, enabling us to refine our intermediate and final outcome measures; it allowed us to test our baseline data collection instruments; and it generated data that we used to decide what sample size we needed for the full-scale evaluation.

The pilot was quite promising. People in the villages were enthusiastic during the small-group discussions and at the large community meetings. There was a high level of attendance and participation in the conversations. Parents tested their children on basic literacy, and people became very engaged. Pratham ran the pilot, and the researchers performed a qualitative analysis of it. Observing the steps that communities took in response to the intervention led to new questions in our survey designed to pick up actions along these lines that other communities might take.

### ***Random assignment***

The evaluation was conducted in 280 villages in Jaunpur District in the state of UP. Districts in India are divided into administrative blocs. In each bloc, on average, there are about 100 villages. Four of these blocs were randomly selected to participate in the study, and the study villages were then randomly selected within each bloc. The study is thus representative of Jaunpur District (and its population of 3.9 million) as a whole.

Because the VECs were seen as key targets of the program and there was only one VEC per village, it was not possible to randomize individual households in and out of the program. An entire village needed to either receive the program (be a treatment village) or be a comparison village. In other words, we needed to randomize at the village level. We worried that the existence of the program in one

village might benefit neighboring villages. For example, one village might complain about school quality to their MP and the MP might then press for changes in all the schools in her constituency. If this happened, it would lead us to under-

estimate the impact of the program. We thought this type of action was unlikely, but we nevertheless decided to track how many complaints were made to MPs. In randomizing villages into different treatment groups and a comparison group we used a technique called stratification. This meant that we were sure to have equal numbers of villages in each treatment group from each block and that the level of reading scores at baseline would be the same for all the different treatment groups and the comparison group. We made the random assignments using a computer-based random number generator to assign each of the 280 villages in our study to one of the four groups described in Table 1.1.

**CHAPTER 4** discusses how to randomize, including whether to randomize at the individual or the group level, how to deal with potential spillovers at the design stage, and whether and how to stratify. It also covers the mechanics of randomization.

**TABLE 1.1** Random assignment of treatment groups for community accountability evaluation

Group	Intervention		
	Participants provided information on how to improve education	Participants create village-based scorecards	Volunteers run Read India camps
Comparison group (85 villages)	—	—	—
Treatment Group 1 (65 villages)	X	—	—
Treatment Group 2 (65 villages)	X	X	—
Treatment Group 3 (65 villages)	X	X	X

Note: An X in a cell indicates that the group receives a given treatment; a dash indicates that the group does not receive that treatment.

### ***Data collection plan***

To plan our data collection we started by mapping exactly how each of the alternative programs being tested could lead to changes in learning levels. Underlying each program alternative was an assumption that people had too little information about the quality of education and that education quality would improve if they participated more in the oversight of schools. The second program variant assumed that being involved in creating information (through the creation of report cards) helps reduce the information gap more efficiently and motivates more action. The Read India program variant assumed that people need a way to respond to poor education quality without going through the government bureaucracy. Figure 1.1 shows a very simplified version of the theory of change for the project.

For each step in the theory of change we developed an indicator. For example, to confirm low schooling quality we measured teacher absenteeism. This involved making surprise visits to schools to see whether teachers were present and teaching.

To test knowledge of learning levels we asked parents how well they thought their children could read and then tested their children. This required that a household survey be administered to a random sample of households in the village. At the same time, we asked about parental involvement in their children's education. (Which schools did their children attend? When did they last visit their children's schools? Did they speak at the village meeting about education? Did they check whether their children attended school or did homework?) We wanted to check whether parents, on hearing about poor learning levels, would take action outside the formal government system—for example, by monitoring homework or sending their children to private schools.

To measure gaps in knowledge of how to influence education quality, we asked parents, village leaders, and members of the VECs if they had heard of the VECs and whether they knew the committees' roles and responsibilities. This required a village leader and VEC member

---

**CHAPTER 5** covers how to use a theory of change to develop a comprehensive data collection plan and how to ensure that the chosen indicators reflect real changes on the ground.

---

survey. We asked these questions again at the end of the project, which allowed us to understand whether the program was successful in reducing knowledge gaps.





Need 	Input 	Output 	Outcome 	Impact
<ul style="list-style-type: none"> <li>• <b>Poor school quality</b> <ul style="list-style-type: none"> <li>- low levels of learning</li> <li>- high rates of teacher absenteeism</li> </ul> </li> <li>• <b>Lack of awareness of</b> <ul style="list-style-type: none"> <li>- learning outcomes</li> <li>- village education committees (VECs)</li> <li>- village funds for education</li> </ul> </li> <li>• <b>Lack of participation</b> <ul style="list-style-type: none"> <li>- by parents in their children's education</li> <li>- by village councils through failure to discuss education at meetings</li> </ul> </li> <li>• <b>Lack of tools to act</b></li> </ul>	<ul style="list-style-type: none"> <li>• <b>Information provided on</b> <ul style="list-style-type: none"> <li>- testing tools</li> <li>- role of VECs</li> <li>- village funds</li> </ul> </li> <li>• <b>Motivation</b> <ul style="list-style-type: none"> <li>- discussions and meetings</li> <li>- encouragement of action</li> </ul> </li> <li>• <b>Training held for</b> <ul style="list-style-type: none"> <li>- Read India program</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>• <b>VECs function</b> <ul style="list-style-type: none"> <li>- Requests for assistant teachers</li> <li>- Monitoring visits of schools</li> </ul> </li> <li>• <b>Parents more involved</b> <ul style="list-style-type: none"> <li>- get extra tutoring</li> <li>- visit schools, talk to teachers</li> <li>- switch to private schools</li> </ul> </li> <li>• <b>Classes held</b> <ul style="list-style-type: none"> <li>- students turn up</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>• <b>Reduced knowledge gap</b></li> <li>• <b>More assistant teachers</b></li> <li>• <b>Reduced teacher absenteeism</b></li> <li>• <b>Additional expenditure on schools</b></li> <li>• <b>Students spend more time on</b> <ul style="list-style-type: none"> <li>- homework</li> <li>- reading</li> </ul> </li> <li>• <b>More students go to private schools</b></li> <li>• <b>Children complete</b> <ul style="list-style-type: none"> <li>- classes</li> </ul> </li> </ul>	<p><b>Sustained improved learning</b></p>

FIGURE 1.1 Simplified theory of change

To measure inputs, the research team (independent of Pratham) carefully monitored implementation of the program, checking that village meetings were held, volunteers trained, and reading camps held.

Our final measure of impact was child learning. At the end of the program we retested the same sample of children on basic literacy and numeracy.

All data collection was carried out by an independent survey company with close monitoring by the research team. A team of monitors, hired directly by the research team, redid a random sample of surveys to check the accuracy of the survey company's work.

### ***Power analysis***

We decided on the number of villages in each treatment group and the comparison group by conducting a power analysis. Because we randomized entire villages in or out of the program, one of the most important determinants of statistical power was the number of villages allocated to each version of the program. However, we also needed to decide how many households to interview per village. We realized that we needed only about 10 households per village to get a reasonably accurate picture of what people knew about the VEC. However, to be able to pick up changes in learning levels we would need data on many more children. We therefore tested all the children in the 10 households we interviewed and also tested children in a further 20 randomly selected households where we did not ask questions of the parents.

We had a strict budget limit that covered both the program and the evaluation. We wanted to test six different versions of the program, but our analysis suggested that we had sufficient statistical power to test only three alternative treatments and a comparison group. We also decided to have more of a sample in our comparison group than in any of our treatment groups. This would enable us to very precisely test the average impact across all the community mobilization approaches compared to the comparison group.

We had to decide how small a difference in learning levels we wanted to be able to detect. We used the results of the previous study on Pratham's urban remedial education program as a guide to what the "success" of the program might look like. The rural program was much less expensive than the urban one because it relied on volunteers. In addition, unlike in the urban program, in the rural program



children could choose to go to extra classes after school, so attendance at the remedial sessions was likely to be much lower, reducing our statistical power. For these reasons we wanted to be able to detect an impact that was smaller than that achieved in Pratham's urban program. We also wanted to be able to separately estimate the impact of the program on children with different initial learning levels—particularly those who started off with very poor reading levels. We therefore did a survey before the program started to be able to identify those children.

The final sample consisted of 2,800 households, 316 schools, 17,533 children (ages 7–14) tested in reading and math, and 1,029 VEC members who were interviewed (including village heads) from the 280 villages.

***Implementing the program (and monitoring that it was implemented well)***

The three interventions were implemented between September 2005 and December 2005. A full-time research team monitored activities in Jaunpur to ensure that the randomization protocol was followed and document how well the program was implemented. The risk that those in the comparison group would benefit from the program was minimized by randomizing at the level of the legal village unit so that individuals in the treatment and comparison groups were geographically separated. Monitoring data suggested not only that the program was properly implemented but also that people responded by attending and speaking up at meetings. All treated villages held at least one meeting, and some held more than one. The meetings were well attended, with good representation and participation from different hamlets, castes, and genders.

In 55 of the 65 villages in Treatment Group 3 (i.e., 84 percent of the total), volunteers started reading camps serving a total of 7,453 children in the villages (135 per village on average). In our random sample of surveyed children, 8 percent had attended the camps in the Treatment Group 3 communities.

The *final (or endline) survey* took place in March and April 2006, three months after the treatment arms had been implemented. Enumerators were urged to make every effort to find the same households

---

**CHAPTER 6** explains how to use power analysis to choose the sample size, the number of treatment arms, and the minimum detectable effect an evaluation wants to detect.

---

---

**CHAPTER 7** explains how to minimize the risk of things going wrong with the evaluation, such as people in the comparison group gaining access to the program or people dropping out of the study.

---

and the same children they had interviewed the previous year. In total, the endline survey included 17,419 children, which included all but 716 of the children in the baseline survey.

### *Analyzing the data*

Once the surveys were complete, the survey company entered all the data into a database in two separate rounds and reconciled any discrepancies between the two versions.

The actual analysis was fairly straightforward for this evaluation. We simply compared the level of knowledge, action, and learning in villages in each treatment group with those in the comparison group, adjusting for the fact that we had randomized at the village level by clustering our standard errors. The difference represented the effect of the program.

The large number of outcomes that could have been affected by the interventions created a danger of “cherry picking” results or “data mining.” In other words, there was a danger of consciously or subconsciously highlighting results that showed large effects and ignoring others. To avoid this risk, we created groups, or “families,” of related variables based on our theory of change and tested them as a group. For example, we grouped all outcomes related to parents’ knowledge of children’s learning levels into one outcome and all outcomes related to parents’ involvement with schools into another. For each family of outcomes we calculated the average effect. Our main outcomes were these average effects across many outcome variables.

Although the community mobilization program worked at a community level, the Read India afterschool camps were attended by some children but not others. In addition to looking at the average effect of

---

**CHAPTER 8** discusses data analysis, including adjusting for how randomization was carried out and for low take-up of the program. It also discusses analysis with multiple outcome variables and the pros and cons of committing in advance to how the data will be analyzed.

---

each intervention, therefore, we also estimated the impact of the Read India program on those children who attended the camps. Rather than looking at outcomes for specific children who attended (which would not be random), we used a technique called “instrumen-

tal variables” by which average treatment effects are adjusted by average take-up rates.

---

## TIMELINE

Our start-to-finish timeline was as follows:

2004	Discussion of which questions were the most important, focusing on community involvement for improving government service delivery
July 2004	Ongoing discussions with Pratham about partnering on a large-scale randomized evaluation of community accountability and their flagship Read India program
July 2004–July 2005	Qualitative fieldwork
March 2005	Selection of villages
March–April 2005	Conduct of census
April 2005	Conduct of baseline survey
April 2005–July 2005	Running of pilot program
September 2005–February 2006	Implementation of the three treatment arms immediately following election of new village leaders
March–May 2006	Conduct of follow-up endline survey
June 2006	Beginning of data analysis and writing up of results
2007 onward	Dissemination of results through discussions with Pratham and Indian policymakers and, more broadly, presentations at the World Bank and academic conferences
2007	Pratham’s receipt of a \$9.1 million grant from the William and Flora Hewlett Foundation and the Bill and Melinda Gates Foundation to help them scale up the Read India program in more than 300 of the 600 districts in India
2010	Publication of academic paper

---

## *What the results of this study mean for policy*

Days before the baseline survey was launched, the research and Pratham teams gathered in Jaunpur to train the enumerators from the survey

company on how to administer Pratham's reading test as part of the survey. Rukmini Banerji, at the time Pratham's head of research and head of the northeastern region, took this moment to speak to the Pratham team. She recounted how they had spent many months developing the new program, field testing it, and learning all the nuances of how to engage the community in conversations and how to get parents invested in their children's learning levels. Now these economists from MIT were going to help evaluate it. "And of course," she said, "they may find that it doesn't work. But if it does not work, we need to know that. We owe it to ourselves and the communities we work with not to waste their and our time and resources on a program that does not help children learn. If we find that this program isn't working, we will go and develop something that will."<sup>14</sup>

Rukmini's words were a reflection of the courage Pratham and many other implementers showed in putting their programs on the line and admitting that they might not work. But her words also summed up the rationale for evaluation.

#### The results of the evaluation

Neither providing information on the channels for improving education nor helping citizens gather information on the status of education in their villages led to greater involvement of parents, VEC members, or teachers in the school system. Nor did these interventions lead to private responses such as extra tutoring or moving children to private schools. Given these results, it is not surprising that there was no impact on learning from the first two interventions. The program helped narrow the knowledge gap (on levels of learning and VEC roles) but only modestly, despite the widespread and enthusiastic participation in community meetings.

In contrast, where Pratham conducted the Read India intervention, not only did volunteers teach almost 7,500 children in after-school camps but literacy rates also improved. The average improvements were modest: a 1.7 percent increase in those who could recognize letters, for example. But the program was designed to help those who could not yet read. Among this group we saw much more impressive effects. Children who could not recognize letters before the program started were 7.9 percent more likely to be able to recognize letters at the end of the program in Treatment Group 3 villages. And once we adjusted for the

14. This quote reflects my memory of Rukmini's speech.

fact that only 13 percent of children who could not recognize letters attended reading camps, we calculated that the camps led to a 60 percent increase in the ability to recognize letters among those who could not recognize letters at baseline. Twenty-six percent of those who could not recognize letters but attended camps could read fluently as a result of the camps.

What did the results imply?

We had two challenges in interpreting the results. We had to understand what had happened in the specific program in UP, why some interventions had worked and others had not. But we also wanted to figure out what this told us about community accountability programs in general. For this we would need to put our results in the context of those from other emerging studies.

In our academic paper we concluded, “In the UP context, providing information on the status of education and the institutions of participation alone is not sufficient to encourage beneficiary involvement in public schools. . . . [However,] information combined with the offer of a direct channel of action can result in collective action and improve outcomes. . . . In the UP context there seemed to be a greater willingness of individuals to help improve the situation for other individuals (via volunteer teaching) rather than collective action to improve institutions and systems.” We noted, “This may be specific to the Indian schooling bureaucracy. Parents may be too pessimistic about their ability to influence the system even if they are willing to take an active role, or parents may not be able to coordinate to exercise enough pressure to influence the system. Nevertheless, the results do suggest that some caution is warranted when recommending standard beneficiary control approaches.”<sup>15</sup>

Pratham responded to the results of the evaluation in a number of ways. Although they did not give up on their objective of changing Indian public schools for the better, they put much less faith in doing it through village councils and VECs. The simple testing tool that was developed in UP is now used to test children throughout India in the Annual State of Education Report. The district and state report cards

15. Abhijit Banerjee, Rukmini Banerji, Esther Duflo, Rachel Glennerster, and Stuti Khemani, “Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India,” *American Economic Journal: Economic Policy* 2 (2010): 1–30, quote on p. 5.

that come out of this testing generate considerable media attention and are intended to put pressure on state and district officials to improve education quality and focus on learning (as opposed to school attendance or school meals). The evaluation results showing the success of Read India helped Pratham win significant additional funding and led to an expansion of Read India to more than 23 million children across India. But Pratham was concerned that although their camps worked for those who attended, only a modest proportion of those who needed help went to the camps. Pratham's long-term goal is to take their pedagogical techniques into India's public school system. They have therefore continued to innovate and evaluate, increasingly working with state governments. Some of their innovations have proved successful, others less so, and continuing to test and evaluate has helped them differentiate between the two outcomes.

At the international level, too, researchers and practitioners have continued to innovate and evaluate how to improve the quality of public services for the poor. A number of studies emerged around the same time as ours. The most similar was a study of a community mobilization program in Uganda that rejuvenated the community oversight committees of local clinics, provided information about the poor quality of services (such as high rates of health worker absenteeism), and worked with communities and health workers to devise ways to improve service quality. The result was reduced absenteeism and improved health.<sup>16</sup> In some ways the results seem in stark contrast to our own, but in other ways there were similarities. In both cases, community participation structures existed prior to the program but were nonfunctional. Arguably, providing communities with very direct actions they could take was a feature of the successful elements of both studies. Since then we have learned that providing information about the quality of schools in Pakistani communities with competition between public and private schools helped improve test scores; empowering school committees in Kenya had no effect,<sup>17</sup> but giving those committees resources to hire

16. Martina Bjorkman and Jakob Svensson, "Power to the People: Evidence from a Randomized Field Experiment on Community-Based Monitoring in Uganda," *Quarterly Journal of Economics* 124 (2009): 735–769.

17. Banerjee et al., "Pitfalls of Participatory Programs"; Christel Versmeersch and Michael Kremer, "School Meals, Educational Achievement and School Competition: Evidence from a Randomized Evaluation," World Bank Policy Research Working Paper 3523, World Bank, Washington, DC, 2004. <http://ssrn.com/abstract=667881> or <http://dx.doi.org/10.2139/ssrn.667881>.

local teachers resulted in higher test scores, and training those committees in their monitoring role enhanced this effect.<sup>18</sup> We also learned that community monitoring of local road projects in Indonesia was less effective than outside audits in reducing corruption.<sup>19</sup>

What can we conclude from all this evidence? We have found that community oversight can improve service quality in the right situations, but it is hard to make it work. There is no single simple answer to the question of whether beneficiary participation works. I was taught in my very first economics lecture that the answer to many economic questions is “It depends,” and in this case the answer seems to depend in complex ways on the details of the program and the institutional setting.

---

**CHAPTER 9** discusses how evidence from randomized evaluations can provide insights for policy, including how to decide when results are likely to generalize and how to make cost-effectiveness comparisons.

---

But we now understand a lot better than we did in 2004 what success is likely to depend on. The accumulated evidence has led to a much more nuanced and informed discussion about community accountability programs and provided considerable food for thought for those designing these programs.

And this is the nature of our journey. We innovate and test. The results move us forward but also generate more questions, which again need to be answered through testing. But over time we learn. We understand more about what is working where and why, and this helps us develop better programs that lead to better lives.

*In this book we seek to give practical advice to those who want to be part of this journey by contributing to the growing base of evidence from randomized evaluations on how to improve the lives of the poor.*

18. This study by Esther Duflo, Pascaline Dupas, and Michael Kremer is summarized as Evaluation 3 in the appendix.

19. Benjamin A. Olken, “Monitoring Corruption: Evidence from a Field Experiment in Indonesia,” *Journal of Political Economy* 115 (2007): 200–249.