3 Use of Randomization in the Evaluation of Development Effectiveness

Esther Duflo and Michael Kremer

Historically, prospective randomized evaluations of development programs have constituted a tiny fraction of all development evaluations. In this chapter, we argue that there is scope for considerably expanding their use, although they must necessarily remain a small fraction of all evaluations.

The benefits of knowing which programs work and which do not extend far beyond any program or agency, and credible impact evaluations are global public goods in the sense that they can offer reliable guidance to international organizations, governments, donors, and nongovernmental organizations (NGOs) beyond national borders. Traditional methods of measuring program impact may be subject to serious bias due to omitted variables.

For a broad class of development programs, randomized evaluations can be used to address these problems. Of course, not all programs can be evaluated with randomized evaluations; for example, examinations of issues such as central bank independence must rely on other methods of evaluation. Programs targeted to individuals or local communities (such as sanitation, local government reforms, education, and health) are likely to be strong candidates for randomized evaluations; this chapter uses the case of educational programs in developing countries as an example.

We do not propose that all projects be subject to randomized evaluations. But we argue that there is currently a tremendous imbalance in evaluation methodology and that increasing the share of projects subject to randomized evaluation from near zero to even a small fraction could have a tremendous impact on knowledge about what works in development. All too often development policy is based on fads, and randomized evaluations could allow it to be based on evidence.

The chapter proceeds as follows. Section 3.1 discusses the methodology of randomized evaluations. We present the impact evaluation problem, review why other current evaluation methods may often be unable to adequately control for selection bias, and discuss why randomized evaluations can be useful in

addressing the problems encountered by other evaluation practices. Section 3.2 reviews recent randomized evaluations of educational programs in developing countries, including programs to increase school participation, provide educational inputs, and reform education. Section 3.3 extracts lessons from these evaluations, and section 3.4 reviews an example of current practice, offers political economy explanations for why randomized evaluations are so rare, and discusses the role international agencies can play in promoting and financing rigorous evaluations, including randomized evaluations. Section 3.5 discusses the value of credible impact evaluations as international public goods.

3.1 The Methodology of Randomized Evaluations

This section discusses the selection bias problem that can arise when conducting impact evaluations and nonrandomized evaluation methods that are used in attempting to control for this bias.

3.1.1 The Evaluation Problem

Any impact evaluation attempts to answer an essentially counterfactual question: How would individuals who participated in the program have fared in the absence of the program? How would those who were not exposed to the program have fared in the presence of the program? The difficulty with these questions is immediate: at a given point in time, an individual is observed to be either exposed or not exposed to the program. Comparing the same individual over time will not, in most cases, give a reliable estimate of the impact the program had on him or her, since many other things may have changed at the same time as the program was introduced. We therefore cannot seek to obtain an estimate of the impact of the program on each individual. All we can hope for is to be able to obtain the average impact of the program on a group of individuals by comparing them to a similar group of individuals who were not exposed to the program.

The critical objective of impact evaluation is therefore to establish a credible comparison group: a group of individuals who in the absence of the program would have had outcomes similar to those who were exposed to the program. This group should give us an idea of what would have happened to the members of the program group if they had not been exposed, and thus allow us to obtain an estimate of the average impact on the group in question.

In reality, however, the individuals who participated in a program generally differ from those who did not. Programs are placed in specific areas (for example, poorer or richer areas), individuals are screened for participation in the

program (for example, on the basis of poverty or on the basis of their motivation), and the decision to participate is often voluntary. For all of these reasons, those who were not exposed to a program are often a poor comparison group for those who were, and any differences between the groups can be attributed to two factors: preexisting differences (the so-called selection bias) and the impact of the program. Since we have no reliable way to estimate the size of the selection bias, we typically cannot decompose the overall difference into a treatment effect and a bias term.

To solve this problem, program evaluations typically need to be carefully planned in advance in order to determine which group is a likely control group. One situation where the selection bias disappears is when the treatment and comparison groups are selected randomly from a potential population of participants (such as individuals, communities, schools, or classrooms). In this case, on average, we can be assured that those who are exposed to the program are no different from those who are not, and thus that a statistically significant difference between the groups in the outcomes the program was planning to affect can be confidently attributed to the program.

As we will see in this chapter, the random selection of treatment and comparison groups can occur in several circumstances. Using the example of PROGRESA, a program designed to increase school participation in Mexico, we discuss how prospective randomized evaluations can be used and how their results can help in scaling successful programs. Using examples of school-based health programs in Kenya and India, we illustrate how prospective randomized evaluations can be used when implementing adapted replications of programs, and using the example of a school voucher program in Colombia, we illustrate how program-induced randomization can occur.

It is worth briefly outlining a few clarifications regarding the use of randomized evaluations to estimate program effects. First, a distinction can be made about what exactly the evaluation is attempting to estimate. Randomized evaluations can be used to estimate the effect of a treatment on either the entire population that was subject to the randomization or on a subset of the population defined by predetermined characteristics, whereas instrumental variable techniques estimate local average treatment effects (Imbens and Angrist 1994; Heckman et al. 1998; Heckman, James, and Todd 1997; Heckman, Lalonde, and Smith 1999). Second, randomized evaluations estimate partial equilibrium treatment effects, which may differ from general equilibrium treatment effects (Heckman, Lochner, and Taber 1998). It is possible that if some educational programs were implemented on a large scale, the programs could affect the functioning of the school system and thus have a different impact.

3.1.2 Other Techniques to Control for Selection and Other Omitted Variable Bias

Natural or organized randomized evaluations are not the only methodologi that can be used to obtain credible impact evaluations of program effect Researchers have developed alternative techniques to control for bias as muc as possible, and progress has been made, most notably by labor economists. Among the techniques that are most popular with researchers are propensit score matching, difference-in-difference estimates, and regression discontinuit design.

One strategy to control for bias is to attempt to find a control group that i as comparable as possible to the treatment group, at least along observabl dimensions. This can be done by collecting as many covariates as possible an then adjusting the computed differences through a regression, or by "matching" the program and the comparison group through forming a comparison group that is as similar as possible to the program group. One possibility is to predict the probability that a given individual is in the comparison or the treatment group on the basis of all available observable characteristics, and then to form a comparison group by picking people who have the same probability of being treated as those who were actually treated (propensity score matching). The challenge with this method, as with regression controls, is that it hinges on having identified all the potentially relevant differences between the treatment and control groups. In cases where the treatment is assigned on the basis of a variable that is not observed by the researcher (demand for the service, for example), this technique can lead to misleading inferences.

A second strategy is what is often called the difference-in-difference technique. When a good argument can be made that the outcome would not have had differential trends in regions that received the program if the program had not been put in place, it is possible to compare the growth in the variables of interest between program and nonprogram regions. However, it is important not to take this assumption for granted. This identification assumption cannot be tested, and even to ascertain its plausibility, one needs to have long time series of data from before the program was implemented in order to be able to compare trends over long enough periods. One also needs to make sure that no other program was implemented at the same time, which is often not the case. Finally, when drawing inferences, one must take into account that regions are often affected by time-persistent shocks that may look like program effects. Bertrand, Duflo, and Mullainathan (2002) found that difference-in-difference estimations (as commonly performed) can severely bias standard errors: the researchers randomly generated placebo laws and found that with about twenty

years of data, difference-in-difference estimates found an "effect" significant at the 5 percent level of up to 45 percent of the placebo laws.

As an example of where difference-in-difference estimates can be used, Duflo (2001) took advantage of a rapid school expansion program that occurred in Indonesia in the 1970s to estimate the impact of building schools on schooling and subsequent wages. Identification was made possible by the fact that the allocation rule for the schools was known (more schools were built in places with low initial enrollment rates) and that the cohorts participating in the program were easily identified (children twelve years or older when the program started did not participate in the program). The increased growth of education across cohorts in regions that received more schools suggests that access to schools contributed to increased education. The trends were quite parallel before the program and shifted clearly for the first cohort that was exposed to the program, thus reinforcing confidence in the identification assumption. However, this identification strategy is not usually valid. Often when policy changes are used to identify the effect of a particular policy, the policy change is itself endogenous to the outcomes it was meant to affect, thus making identification impossible (Besley and Case 2000).

Finally, a third strategy, called regression discontinuity design (Campbell 1969), takes advantage of the fact that program rules sometimes generate discontinuities that can be used to identify the effect of the program by comparing those above a certain threshold to those just below it. If resources are allocated on the basis of a certain number of points, it is possible to compare those just above to those just below the threshold. Angrist and Lavy (1999) use this technique to evaluate the impact of class size in Israel, where a second teacher is allocated every time the class size grows above forty students. This policy generates discontinuities in class size when the enrollment in a grade grows from forty to forty-one (as class size changes from one class of forty students to one class each of twenty and twenty-one students). Angrist and Lavy compared test scores in classes just above and just below this threshold and found that those just above the threshold have significantly higher test scores than those just below, which can confidently be attributed to the class size, since it is very unlikely that schools on both sides of the threshold have any other systematic differences.[2] Such discontinuities in program rules, when enforced, are thus sources of identification.

In developing countries, however, it is often likely to be the case that rules are not enforced strictly enough to generate discontinuities that can be used for identification purposes. For example, researchers attempted to use as a source of identification the discontinuity in the policy of the Grameen bank (the flagship microcredit organization in Bangladesh), which is to lend only to

people who own less than one acre of land (Pitt and Khandker 1998). It turns out that in practice, the Grameen bank lends to many people who own more than one acre of land and that there is no discontinuity in the probability of borrowing at the threshold (Morduch 1998).

These three techniques are subject to large biases that can lead to either overestimation or underestimation of program impact. LaLonde (1986) found that many of the econometric procedures and comparison groups used in program evaluations did not yield accurate or precise estimates and that such econometric estimates often differ significantly from experimental results.

Identification issues with nonrandomized evaluation methods must be tackled with extreme care because they are less transparent and more subject to divergence of opinion than are issues with randomized evaluations. Moreover, the differences between good and bad nonrandomized evaluations are difficult to communicate, especially to policymakers, because of all the caveats that must accompany the results. In practice, these caveats may never be provided to policymakers, and even if they are provided, they may be ignored. In either case, policymakers are likely to be radically misled. This suggests that while nonrandomized evaluations will continue to be needed, there should be a commitment to conduct randomized evaluations where possible.

3.2 Examples of Randomized Evaluations of Educational Programs

In this section, we present recent randomized evaluations of three types of educational programs in developing countries: programs designed to increase school participation, programs providing educational inputs, and educational reform programs.

3.2.1 Increasing School Participation

Education is widely considered to be critical for development: the internationally agreed-on Millennium Development Goals call for universal primary school enrollment by 2015. However, there is considerable controversy over how best to achieve this goal and how much it would cost. For example, some argue that it will be difficult to attract additional children to school since most children who are not in school are earning income their families need, while others argue that children of primary-school age are not very productive and that modest incentives or improvements in school quality would be sufficient. Some see school fees as essential for ensuring accountability in schools and as a minor barrier to participation, while others argue that eliminating fees would greatly increase school participation.

Because one obvious means of increasing school participation is to decrease or remove financial barriers, we review recent randomized evaluations of programs designed to increase school participation through reducing the cost of school, or even paying for school attendance.[3]

PROGRESA

Because positive results can help to build a consensus for a project, carefully constructed program evaluations form a sound basis for decisions on whether to scale up existing projects. The PROGRESA program in Mexico, designed to increase school participation, is a striking example of this phenomenon. PROGRESA provides cash grants to women that are conditional on children's school attendance and preventative health measures (nutrition supplementation, health care visits, and participation in health education programs). When the program was launched in 1998, officials in the Mexican government made a conscious decision to take advantage of the fact that budgetary constraints made it impossible to reach the 50,000 potential participant communities of PROGRESA immediately, and instead began with a program in 506 communities. Half of those communities were randomly selected to receive the program, and baseline and subsequent data were collected in the remaining communities (Gertler and Boyce 2001). Part of the rationale for this decision was to increase the probability that the program would be continued if there were a change in the party in power, because the proponents of the program understood that the program would require continuous political support in order to be scaled up successfully. The task of evaluating the program was given to academic researchers through the International Food Policy Research Institute (IFPRI); the data were made accessible to numerous researchers, and a number of papers have been written on PROGRESA's impact.[4]

The evaluations show that the program was effective in improving both health and education. Comparing PROGRESA participants and nonparticipants, Gertler and Boyce (2001) show that children on average had a 23 percent reduction in the incidence of illness, a 1 to 4 percent increase in height, and an 18 percent reduction in anemia. Adults experienced a reduction of 19 percent in the number of days lost due to illness. Shultz (2004) finds an average 3.4 percent increase in enrollment for all students in grades 1 through 8; the increase was largest among girls who had completed grade 6, at 14.8 percent.

In part because the randomized phase-in of the program allowed such clear documentation of the program's positive effects, PROGRESA was indeed maintained when the Mexican government changed hands: by 2000, PROGRESA was reaching 2.6 million families (10 percent of the families in Mexico) and had a budget of US$800 million, or 0.2 percent of gross domestic

product (GDP) (Gertler and Boyce 2001). The program was subsequently expanded to urban communities, and with support from the World Bank, similar programs are being implemented in several neighboring Latin American countries. Mexican officials transformed a budgetary constraint into an opportunity and made evaluation the cornerstone of subsequent scaling up. They were rewarded by both the expansion of the program and the tremendous visibility that the program acquired.

School Meals, Cost of Education, and School Health in Kenya: Comparing the Cost-Effectiveness of Different Interventions

A central policy concern for developing countries is the relative cost-effectiveness of various interventions intended to increase school participation. This section discusses research on several programs to decrease the costs of education and compares the cost-effectiveness of these different interventions.

Evaluations of cost-effectiveness require knowledge of a program's costs as well as its impact, and comparability across studies requires some common environment. It is difficult to compare the impact of PROGRESA's cash transfers with that of, say, school meals in Kenya, since it is unclear whether the resulting differences are associated with the type of program or the larger environment. In general, analysts and policymakers are left with a choice between retrospective studies, which allow comparison of different factors affecting school participation, and randomized evaluations, which yield very credible estimates of the effect of single programs. One exception to our general inability to compare cost-effectiveness estimates is a recent set of studies conducted in Kenya of programs seeking to improve school participation. By evaluating a number of programs in a similar setting (a specific district in Western Kenya), it is possible to explicitly compare the cost-effectiveness of different approaches to increasing school participation. Looking at the effect of school meals on school participation, Vermeersch (2002) found that school participation was 30 percent greater in twenty-five Kenyan preschools where a free breakfast was introduced than it was in twenty-five comparison schools. However, the provision of meals cut into instruction time. Overall, test scores were .4 standard deviations greater in the program schools, but only if the teacher was well trained prior to the program.

Kremer, Moulin, and Namunyu (2002) evaluate a program in which a nongovernmental organization, Internationaal Christelijk Steunfonds Africa, provided uniforms, textbooks, and classroom construction to seven schools that were randomly selected from a pool of fourteen poorly performing candidate schools in Kenya. As in many other countries, parents face significant private costs of education for school fees and other expenses, such as uniforms. In par-

ticular, they are normally required to purchase uniforms at about $6, a substantial expense in a country with per capita income of $340. Dropout rates fell considerably in treatment schools, and after five years, pupils in treatment schools had completed about 15 percent more schooling. In addition, many students from nearby schools transferred into program schools, raising class size by 50 percent. This suggests that students and parents were willing to trade off substantially larger class sizes for the benefit of free uniforms, textbooks, and improved classrooms. Given that the combination of these extra inputs and a 50 percent increase in class size led to no measurable impact on test scores, but that the cost savings from a much smaller increase in class size would have allowed the Kenyan government to pay for the uniforms, textbooks, and other inputs provided under the program, these results suggest that existing budgets could be productively reallocated to decrease parental payments and substantially increase school participation.

Poor health may also limit school participation: for example, intestinal helminthes such as hookworm affect a quarter of the world's population and are particularly prevalent among school-age children. Miguel and Kremer (2004) evaluate a program of twice-yearly school-based mass treatment with inexpensive deworming drugs in Kenya, where the prevalence of intestinal worms among children is very high. Seventy-five schools were phased into the program in random order. Health and school participation improved not only at program schools but also at nearby schools due to reduced disease transmission. Absenteeism in treatment schools was 25 percent (or seven percentage points) lower than in comparison schools. Including the spillover effect, the program increased schooling by 0.15 years per person treated.

Because these programs were conducted in similar environments, cost-effectiveness estimates from numerous randomized evaluations can be readily compared. Deworming was found to be extraordinarily cost-effective at only $3.50 per additional year of schooling (Miguel and Kremer 2003). In contrast, even under optimistic assumptions the provision of free uniforms would cost $99 per additional year of school participation induced (Kremer, Moulin, and Namunyu 2002). The school meals program, which targeted preschoolers rather than primary school age children, cost $36 per additional year of schooling induced (Vermeersch 2003). This suggests that school health programs may be one of the most cost-effective ways of increasing school participation.

3.2.2 School Inputs

This section reviews recent randomized evaluations of programs that provide various inputs to schools in Kenya and India.

Retrospective and Prospective Studies of Inputs in Kenyan Primary Schools

Based on existing retrospective evaluations, many are skeptical about th effects of educational inputs on learning (Hanushek 1995). One potential weak ness of such evaluations is that observed inputs may be correlated with omitte variables that affect educational outcomes. The evaluation could be biased up ward, for example, if observed inputs are correlated with unobserved parenta or community support for education, or downward if compensatory program provide assistance to poorly performing schools.

Although retrospective studies provide at best mixed evidence on the effec of many types of school inputs, they typically suggest that the provision o additional textbooks in schools with low initial stocks can improve learning Indeed, cross-sectional and difference-in-difference analyses of Kenyan dat would suggest that textbooks have dramatic effects on test scores. Result from a randomized evaluation, however, point to a subtler picture. Provisio of textbooks increased test scores by about 0.2 standard deviations, but onl among students who had scored in the top one or two quintiles on pretest prior to the program. Textbook provision did not affect the scores of the bot tom 60 percent of students (Glewwe, Kramer, and Moulin 2002). Many stu dents may have failed to benefit from textbooks because they had difficult understanding them: Kenyan textbooks are in English, the official language o instruction, but English is most pupils' third language, after their mothe tongue and Swahili. More generally, the Kenyan curriculum is set at a leve that, while perhaps appropriate for elite families in Nairobi, is far ahead o that typically attained by rural students, given the high rates of student an teacher absence from school.

Given the results of the textbook study, researchers tried providing fli charts, an alternative input that presumably was more accessible to weal pupils. Glewwe and others (2004) compared retrospective and prospective anal yses of the effect of flip charts on test scores. Retrospective estimates usin straightforward ordinary-least-squares regressions suggest that flip charts rais test scores by up to 20 percent of a standard deviation, robust to the inclusio of control variables. Difference-in-difference estimates suggest a smaller effec of about 5 percent of a standard deviation, an effect that is still significant though sometimes only at the 10 percent level. In contrast, prospective esti mates based on randomized evaluations provide no evidence that flip chart increase test scores. These results suggest that using retrospective data to com pare test scores seriously overestimates the charts' effectiveness. A difference-in difference approach reduced but did not eliminate this problem. Moreover, it i not clear that such a difference-in-difference approach has general applicability

These examples suggest that the ordinary-least-squares estimates are biased upward rather than downward. This is plausible, since in a poor country with a substantial local role in education, inputs are likely to be correlated with favorable unobserved community characteristics. If the direction of omitted variable bias were similar in other retrospective analyses of educational inputs in developing countries, the effects of inputs may be even more modest than retrospective studies suggest.

Placing Additional Teachers in Nonformal Education Centers

Banerjee and others (2001) evaluated a program in which Seva Mandir, an Indian NGO, placed second teachers in nonformal education centers that the NGO runs in Indian villages. These nonformal schools seek to provide basic numeracy and literacy skills to children who do not attend formal school and, in the medium term, to help mainstream these children into the regular school system. The centers are plagued by high teacher and child absenteeism. A second teacher (when possible, a woman) was randomly assigned to twenty-one out of forty-two of these centers, and the hope was to increase the number of days the centers were open, increase children's participation, and increase performance by providing more individualized attention to the children. By providing a female teacher, the NGO also hoped to make school more attractive for girls. Teacher attendance and child attendance were regularly monitored throughout the duration of the project.

The project reduced the number of days a center was closed: one-teacher centers were closed 44 percent of the time, whereas two-teacher centers were closed only 39 percent of the time. Girls' attendance had increased by 50 percent. However, there were no differences in test scores. It is worth noting that careful evaluations form a sound basis for decisions of whether to scale up existing projects. In the example, the two-teacher program was not implemented on a full scale by the NGO, on the grounds that the benefits were not sufficient to outweigh the cost. The savings were used to expand other programs.

Remedial Education Programs

Pratham, an Indian NGO, implemented a remedial education program in 1994 that now reaches more than 161,000 children in twenty cities. The program hires young women from the communities to provide remedial education in government schools to children who have reached grade 2, 3, or 4 without having mastered the basic grade 1 competencies. Children who are identified as lagging are pulled out of the regular classroom for two hours a day to receive this instruction. Pratham wanted to evaluate the impact of this program, one of the NGO's flagship interventions, at the same time as it was looking to

expand it; the expansion into a new city, Vadodara, provided an opportunity to conduct a randomized evaluation (Banerjee et al. 2003). In the first year (1999–2000), the program was expanded to 49 (randomly selected) of the 123 Vadodara government schools. In 2000–2001, the program was expanded to all the schools, but half the schools received a remedial teacher for grade 3 and half received one for grade 4. Grade 3 students in schools that were exposed to the program in grade 4 serve as the comparison group for grade 3 students who were directly exposed to the program. Simultaneously, a similar intervention was conducted in a district of Mumbai, where half the schools received the remedial teachers in grade 2 and half received the teachers in grade 3. The program was continued for an additional year, with the school switching groups.

The program was thus conducted in several grades, in two cities, and with all schools participating in the program. On average, after two years, the program increased student test scores by 0.39 standard deviations. Moreover, the gains were largest for children at the bottom of the distribution: children in the bottom third gained 0.6 standard deviations after two years. The impact of the program is rising over time and is very similar across cities and child gender. Hiring remedial education teachers from the community appears to be ten times more cost-effective than hiring new teachers. One can be relatively confident in recommending the scaling up of this program, at least in India, on the basis of these estimates, since the program was continued for a period of time, was evaluated in two very different contexts, and has shown its ability to be rolled out on a large scale.

3.2.3 School Reform

There is reason to believe that many school systems could benefit from considerable reform. For example, evidence from the Kenyan evaluations discussed previously suggests that budgets are misallocated and that the curriculum focuses excessively on the strongest students. Teacher incentives in Kenya, as in much of the rest of the developing world, are quite weak, and absence among teachers is quite high, at around 20 percent. Proposed school reforms range from decentralization of budget authority to strengthening links between teacher pay and performance to vouchers and school choice. As an example, a decentralization program in Kenya that provided small grants to parent-run school committees induced them to purchase textbooks, with educational consequences similar to those of the textbook program mentioned above (Glewwe, Ilias, and Kremer 2003). Providing larger grants led school committees to shift their spending toward construction, but no educational impact could be observed from this, at least in the short run.

Teacher Incentives

Some parent-run school committees in Kenya provide gifts to teachers whose students perform well. Glewwe, Ilias, and Kremer (2003) evaluate a program that provided prizes to teachers in schools that performed well on exams and had low dropout rates. In theory, this type of incentive could lead teachers to either increase effort or, alternatively, teach to the test. Empirically, teachers responded to the program by teaching to the test: they did not increase their attendance but provided more sessions to prepare students for the exams. Consistent with a model in which teachers respond by increasing their effort to manipulate test scores rather than to stimulate long-term learning, the test scores of students who had been part of the program initially increased, but by the end of the program, they had fallen back to levels similar to those of the comparison group.

School Vouchers

Angrist and others (2002) evaluate a Colombian program in which vouchers for private schools were allocated by lottery because of limitations in the program's budget. Vouchers were renewable, conditional on satisfactory academic performance. The researchers found that lottery winners were 15 to 20 percent more likely to attend private school and 10 percent more likely to complete eighth grade, and they scored 0.2 standard deviations higher on standardized tests, equivalent to a full grade level. The effects of the program were greater for girls than for boys. Winners were substantially more likely to graduate from high school, and they scored higher on high school completion and college entrance exams. The benefits of the program to participants clearly exceeded the additional cost relative to the alternative of providing places in public schools.

3.3 Lessons

The evaluations we have described offer both substantive and methodological lessons. School participation can be substantially increased through implementing inexpensive health programs, reducing the costs of school to households, or providing school meals. Given the features of the education system in Kenya, which like many other developing countries has a curriculum focused on the strongest students, limited teacher incentives, and suboptimal budget allocation, simply providing more resources may have a limited impact on school quality. A remedial education program in India suggests that it is possible to improve student test scores substantially at a very low cost. Decentralizing budgets to school committees or providing teacher incentives based on test

scores had little impact in Kenya, but a school choice program in Colombia yielded dramatic benefits for participants.

We review some of the methodological lessons that can be drawn from the examples discussed.

3.3.1 Results from Randomized Evaluations Can Be Quite Different from those Drawn from Retrospective Evaluations

As seen in the studies of textbooks and flip charts in Kenya, estimates from prospective randomized evaluations can often be quite different from the effects estimated in a retrospective framework, suggesting that omitted-variable bias is a serious concern (Glewwe, Ilias, and Kremer 2003). Similar disparities between retrospective and prospective randomized estimates arise in studies of the impact of deworming in Kenya (Miguel and Kremer 2003a) and the impact of social networks on the take-up of deworming drugs (Miguel and Kremer 2003b).

Comparative studies that estimate a program's impact using experimental methods and then reestimate impact using one or several different nonexperimental methods suggest that omitted-variable bias is a significant problem beyond just the examples mentioned here. Although we are not aware of any systematic review of studies in developing countries, one recent study in developed countries suggests that omitted-variable bias is a major problem when nonexperimental methods are used (Glazerman, Levy, and Meyers 2002). This study assessed both experimental and nonexperimental methods in the context of welfare, job training, and employment service programs and found that nonexperimental estimators often produce results dramatically different from those of randomized evaluations, that the estimated bias is often large, and that no strategy seems to perform consistently well.[5]

Future research along these lines would be valuable, as such comparative studies can help to show the extent to which the biases of retrospective estimates are significant. However, when the comparison group for the nonexperimental portions of these comparative studies is decided ex post, the evaluator may be able to pick from a variety of plausible comparison groups, some of which may have results that match experimental estimates and some of which may not. (This is also an issue for retrospective studies in regard to problems with publication bias.) Possible ways of addressing these concerns in the future include conducting nonexperimental evaluations first, before the results of randomized evaluations are released, or having researchers conduct blind nonexperimental evaluations without knowledge of the results of randomized evaluations or other nonexperimental studies.

3.3.2 Randomized Evaluations Are Often Feasible

As is clear from the examples discussed in this chapter, randomized evaluations are feasible and have been conducted successfully. They are labor intensive and costly, but no more so than other data collection activities. Political economy concerns may sometimes make it difficult not to implement a program in the entire population. For example, Opportunidades, the urban version of PROGRESA, will not start with a randomized evaluation because of the strong opposition to delaying access to the program. Such concerns can be tackled at several levels. For example, when financial or administrative constraints necessitate phasing in programs over time, randomization may be the fairest way of determining the order of phase-in.

3.3.3 NGOs Are Well Suited to Conduct Randomized Evaluations, But Will Require Technical Assistance (for example, from Academics) and Outside Financing

Governments are not the only vehicles through which randomized evaluations can be organized. Indeed, the evidence presented in this chapter suggests that one possible model is that of evaluation of NGO projects. Unlike governments, NGOs are not expected to serve entire populations. Even small NGOs can substantially affect budgets in developing countries. Given that many NGOs exist and that they frequently seek out new projects, it is often relatively straightforward to find NGOs willing to conduct randomized evaluations; hitches are more often logistical than philosophical.

For example, the set of recent studies conducted in Kenya has been carried out through a collaboration with the Kenyan NGO Internationaal Christelijk Steunfonds (ICS) Africa. ICS was keenly interested in using randomized evaluations to see the impact its programs are having, as well in sharing credible evaluation results with other stakeholders and policymakers. A second example is the collaboration between the Indian NGO Pratham and MIT researchers, which led to the evaluations of the remedial education and computer-assisted learning programs (Banerjee et al. 2003). This collaboration was initiated when Pratham was seeking partners to evaluate its programs. Pratham understood the value of randomization and was able to convey the importance of such evaluations to the teachers involved in the project.

However, although NGOs are well placed to conduct randomized evaluations, it is less reasonable to expect them to finance these evaluations. The evaluations of the ICS deworming programs were made possible by financial support from the World Bank, the Partnership for Child Development, U.S. National Institutes of Health, and the MacArthur Foundation. In the case of

the Indian educational programs, Pratham was able to find a corporate spo sor; India's second-largest bank, ICICI Bank, was keenly interested in evalua ing the impact of the program and helped to finance part of the evaluation. I general, given that accurate estimates of program effects are international pul lic goods, randomized evaluations should be financed internationally.

3.3.4 Costs Can Be Reduced and Comparability Enhanced by Conducting a Series of Evaluations in the Same Area

Once staff are trained, they can work on multiple projects. Since data collectic is the most costly element of these evaluations, cross-cutting the sample ca also dramatically reduce costs. For example, many of the programs seeking t increase school participation were implemented in the same area and by th same organization. The teacher incentives (Glewwe, Ilias, and Kremer 200: and textbook (Kremer, Moulin, and Namunyu 2002) programs were evaluate in the same one hundred schools: one group had textbooks only, one had tex books and incentives, one had incentives only, and one had neither. The effe of the incentive program should thus be interpreted as the effect of an incenti program conditional on half the schools having extra textbooks. In India, computer-assisted learning program was implemented in Vadodara in th same set of schools as the remedial education study.

This tactic must take into account potential interactions between progran (which can be estimated if the sample is large enough) and may not be appr priate if one program makes the schools atypical.

3.3.5 Randomized Evaluations Have a Number of Limitations, But Many of These Limitations Also Apply to Other Techniques

Many of the limitations of randomized evaluations also apply to other tec niques. Here we review four issues that affect both randomized and no randomized evaluations (sample selection bias, attrition bias, spillover effect and behavioral responses) and argue that randomized methods often allow ea ier correction for these limitations than do nonrandomized methods.

Sample selection problems could arise if factors other than random assig ment influence program allocation. For example, parents may move their chi dren out of a school without the program into a school with the prograr Conversely, individuals allocated to a treatment group may not receive th treatment (for example, because they decide not to take up the program Even if randomized methods have been used and the intended allocation the program was random, the actual allocation may not be. This problem ca be addressed through intention to treat (ITT) methods or by using random a signment as an instrument of variables for actual assignment. Although the ir

tial assignment does not guarantee in this case that someone is actually in either the program or the comparison group, in most cases it is at least more likely that someone is in the program group if he or she was initially allocated to it. The researcher can thus compare outcomes in the initially assigned group and scale up the difference by dividing it by the difference in the probability of receiving the treatment in those two groups, to obtain the local average treatment effect estimate (Imbens and Angrist 1994). Methods such as ITT estimates allow selection problems to be addressed fairly easily in the context of randomized evaluations, but it is often much more difficult to make these corrections in the case of a retrospective analysis.

A second issue affecting both randomized and nonrandomized evaluations is differential attrition in the treatment and the comparison groups: those who participate in the program may be less likely to move or otherwise drop out of the sample than those who do not. For example, the two-teacher program analyzed by Banerjee and others (2001) increased school attendance and reduced dropout rates. This means that when a test was administered in the schools, more children were present in the program schools than in the comparison schools. If children who are prevented from dropping out by the program are the weakest in the class, the comparison between the test scores of children in treatment and control schools may be biased downward. Statistical techniques can be used to bound the potential bias, but the ideal is to try to limit attrition as much as possible. For example, in the evaluation of the remedial education program in India (Banerjee et al. 2003), an attempt was made to track down all children and administer the test to them even if they had dropped out of school. Only children who had left for their home village were not tested. As a result, the attrition rate remained relatively high but did not differ between the treatment and comparison schools, increasing confidence in the estimates.

Third, programs may create spillover effects on people who have themselves not been treated. These spillovers may be physical, as found for the Kenyan deworming program by Miguel and Kremer (2003a) when deworming interferes with disease transmission and thus reduces worm infection among both children in the program schools who did not receive the medicine and children in neighboring schools. Such spillovers might also operate through prices, as when the provision of school meals leads competing local schools to reduce school fees (Vermeersch 2002).

Finally, there might also be learning and imitation effects (Duflo and Saez 2003; Miguel and Kremer 2003b).

If such spillovers are global (for example, due to changes in world prices), total program impacts will be difficult to identify with any methodology. However, if such spillovers are local, then randomization at the level of groups

can allow estimation of the total program effect within groups and can gener sufficient variation in local treatment density to measure spillovers acr groups. For example, the solution in the case of the deworming study was choose the school (rather than the pupils within a school) as the unit of r domization (Miguel and Kremer 2003a) and to look at the number of tre ment and comparison schools within neighborhoods. Of course, this require larger sample size.

One issue that may not be as easily dealt with is that the provision of inp might temporarily increase morale among students and teachers, and hence i prove performance. While this would bias randomized evaluations, it wo also bias fixed-effect or difference-in-difference estimates. However, it is uncl how serious an issue this is in practice, whereas we know that selection is a rious concern.

In summary, while randomized evaluation is not a bullet-proof strategy, t potential for biases is well known and can often be corrected. This stands contrast to biases of most other types of studies, where the bias due to the nc random treatment assignments often cannot be signed nor estimated.

3.3.6 Publication Bias Appears to Be Substantial with Retrospective Studies; Randomized Evaluations Can Help Address Publication Bias Problems, But Institutions Are Also Needed

Publication bias is a particularly important issue that must be addressed. Po tive results naturally tend to receive a large amount of publicity: agencies th implement programs seek publicity for their successful projects, and academi are much more interested in and able to publish positive results than modest insignificant results. However, clearly many programs fail, and publication bi will be substantial if positive results are much more likely to be publishe Available evidence suggests the publication bias problem is severe (DeLo and Lang 1992) and especially significant with studies that employ nonexpe mental methods.

Publication bias is likely to be a particular problem with retrospective stu ies. Ex post, the researchers or evaluators define their own comparison gro and thus may be able to pick a variety of plausible comparison groups; in pa ticular, researchers obtaining negative results with retrospective techniques a likely to try different approaches or not to publish. In the case of "natur experiments" and instrumental variable estimates, publication bias may act ally more than compensate for the reduction in bias caused by the use of an i strumental variable, because these estimates tend to have larger standard erro and because researchers looking for significant results will select only large est mates. For example, Ashenfelter, Harmon, and Oosterbeek (1999) show th

there is strong evidence of publication bias in instrumental variables-based estimates of the returns to education: on average, the estimates with larger standard errors also tend to be larger. This accounts for most of the oft-cited result that instrumental estimates of the returns to education are higher than ordinary-least-squares estimates.

In contrast, randomized evaluations commit in advance to a particular comparison group: once the work is done to conduct a prospective randomized evaluation, the results are usually documented and published even if the results suggest quite modest effects or even no effects at all.

As we will discuss in the next section, it is important to put institutions in place to ensure negative results are disseminated. Such a system is already in place for medical trial results, and creating a similar system for documenting evaluations of social programs would help to alleviate the problem of publication bias. Beyond presenting a clearer picture of which interventions have worked and which have not, this type of institution would provide the level of transparency necessary for systematic literature reviews to be less biased in their conclusions about the efficacy of particular policies and programs.

3.3.7 Although Any Given Randomized Evaluation Is Conducted within a Specific Framework with Unique Circumstances, Randomized Evaluations Can Shed Light on General Issues

Without a theory of why a program has the effect it has, generalizing from one well-executed randomized evaluation may be unwarranted. But similar issues of generalizability arise no matter what evaluation technique is being used. One way to learn about generalizability is to encourage adapted replications of randomized evaluations in key domains of interest in several different settings. It will always be possible that a program that failed in one context would have succeeded in another, but adapted replications, guided by a theory of why the program was effective, will go a long way toward alleviating this concern. This is one area where international organizations, already present in most countries, can play a key role. Such an opportunity was seized in implementing adapted replications of PROGRESA in other Latin American countries. Encouraged by the success of PROGRESA in Mexico, the World Bank encouraged (and financed) Mexico's neighbors to adopt similar programs. Some of these programs have included randomized evaluations (for example, the Programa de Asignación Familiar program in Honduras) and are being evaluated.

Often the results of the first phase of a project may be difficult to interpret because of circumstances that are unique to that phase: a project may have failed as the result of implementation problems that could be avoided later in

the project, or a project may have succeeded because it received more resour than a project in a more realistic situation or less favorable context. Even if t choice of the comparison and treatment groups ensures the internal validity estimates, any method of evaluation is subject to problems with external val ity due to the specific circumstances of implementation. That is, the results m not be able to be generalized to other contexts.

One problem specific to randomized evaluations is that members of eitl the treatment or comparison group could potentially change their behavi due not to the intervention but simply to the fact that they would know tl they are a part of a randomized evaluation. Of course, to the extent that bc groups change their behavior in the same way, this will not lead to bias. It also perhaps less likely that this will occur immediately after the introducti of the intervention and over a long period.

One way to address questions about the external validity of any particul study, whether it is a randomized evaluation or not, is to implement adapt replications of successful (and potentially unsuccessful) programs in differe contexts. Such adapted replications have two advantages: first, in the proc of transplanting a program, circumstances will change and robust prograi will show their effectiveness by surviving these changes; second, obtaining se eral estimates in different contexts will provide some information about whet er the program has notably different impacts on different groups. Replicati of the initial phase of a study in a new context does not imply delaying fu scale implementation of the program if that is justified on the basis of existi knowledge. More often than not, however, the introduction of the program c proceed only in stages, and the evaluation requires only that participants phased into the program in random order. In addition, such adapted replic tions can be used to check whether program effects within samples vary wi covariance. For example, suppose that the effect of a given program is small in schools with good teachers; one might consider whether in a different setti with much better teachers, the effect would be smaller.

One example is the work in India of Bobonis, Miguel, and Sharma (200 who conducted an adapted replication of the deworming study in Kenya. T baseline revealed that although worm infection was present, the levels of infe tion were substantially lower than in Kenya (in the India case, "only" 27 pe cent of children suffered from some form of worm infection). However, ' percent of children had moderate to severe anemia, and thus the program w modified to include iron supplementation. The program was administer through a network of preschools in urban India. After a year of providi treatment, the researchers found a nearly 50 percent reduction in moderate severe anemia, large weight gains, and a 7 percent reduction in absenteeis

among four to six year olds (though not for younger children). Their findings support the conclusion of the deworming research in Kenya (Miguel and Kremer 2003a) that school health programs may be one of the most cost-effective ways of increasing school participation and, importantly, suggest that this conclusion may be relevant in low-income countries outside Africa.

It is worth noting that the exogenous variation created by randomization can be used to help identify a structural model. Attanasio, Meghir, and Santiago (2001) and Berhman, Sengupta, and Todd (2002) are two examples of using this exercise in combination with the PROGRESA data to predict possible effects of varying the schedule of transfers. For example, Attanasio, Meghir, and Santiago (2001) found that the randomized component of the PROGRESA data induced extremely useful exogenous variation that helped in the identification of a richer and more flexible structural model. These studies rest on assumptions that one is free to believe or not, but at least they are freed of some assumptions by the presence of this exogenous variation.

The more general point is that randomized evaluations do not preclude the use of theory or assumptions: in fact, they generate data and variation that can be useful in identifying some aspects of these theories. For example, evaluations suggest that the Kenyan educational system is heavily geared toward top students and that reallocating budgets within primary education could lead to considerably better outcomes, pointing to perverse incentives created by Kenya's mix of local and national school finance (see Kremer, Moulin, and Namunyu 2002; Glewwe, Kremer, and Moulin 2002).

3.4 The Role International Agencies Can Play

In this section we review an example of current practice that failed to provide opportunities for rigorous evaluations due to a lack of planning, then present some political economy arguments for why randomized evaluations are so rare, and finally discuss how international agencies can support the use of credible evaluation methods, including randomized evaluations.

3.4.1 The District Primary Education Program: An Example of Lost Opportunity

The District Primary Education Program (DPEP) in India, the largest World Bank–sponsored education program, is an example of a large program with potentially interesting evaluations that have been jeopardized by lack of planning.[6] DPEP was meant to be a showcase example of the ability to go to scale with education reform (Pandey 2000). It is a comprehensive program involving

teacher training, inputs, and classrooms that seeks to improve the performa of public education. Districts are generally given a high level of discretio how to spend the additional resources.

Despite the apparent commitment to a careful evaluation of the progr several features make a convincing impact evaluation of DPEP impossi First, the districts were selected according to two criteria: low level of achi ment (as measured by low female literacy rates) and high potential for impr ment. In particular, the first districts chosen to receive the program w selected "on the basis of their ability to show success in a reasonable t frame" (Pandey 2000, quoted in Case 2001). The combination of these two ments in the selection process makes clear that any comparison between level of achievement of DPEP districts and non-DPEP districts would proba be biased downward, while any comparison between improvement of achi ment between DPEP and non-DPEP districts (difference-in-difference) wo probably be biased upward. This has not prevented the DPEP from putt enormous emphasis on monitoring and evaluation: large amounts of d were collected and numerous reports commissioned. However, the data col tion process was conducted only in DPEP districts. These data will be us only for before-and-after comparisons, which clearly do not make sense in economy undergoing rapid growth and transformation. A researcher w found a credible identification strategy would have to use existing data, s as census or National Sample Survey data.

3.4.2 Why Are Randomized Evaluations So Rare? Some Political Economy Arguments

We have argued that the problems of omitted-variable bias that randomi evaluations are designed to address are real and that randomized evaluati are feasible. They are no more costly than other types of surveys and are cheaper than pursuing ineffective policies. So why are they so rare? C (2001) attributes their rarity in education to the postmodern culture in Am can education schools, which is hostile to the traditional conception of cau tion that underlies statistical implementation. Pritchett (2002) argues t program advocates systematically mislead swing voters into believing exagg ated estimates of program impacts. Advocates block randomized evaluati since they would reveal programs' true impacts to voters.

A complementary explanation is that policymakers are not systematic fooled but rather have difficulty gauging the quality of evidence in part beca advocates can suppress unfavorable evaluation results. Suppose retrospect regressions yield estimated program effects equal to the true effect plus m surement error plus a bias term, possibly with mean of zero. Program ad

cates then select the highest estimates to present to policymakers and any opponents select the most negative estimates. Knowing this, policymakers rationally discount these estimates. For example, if advocates present a study showing 100 percent rate of return, the policymaker might assume the true return is 10 percent. In this environment there is little incentive to conduct randomized evaluations. Since the resulting estimates include no bias term, they are unlikely to be high enough or low enough that advocates will present them to policymakers. Even if results are presented to policymakers, those policymakers unable to gauge the quality of particular studies will discount them. Why fund a project that a randomized evaluation suggests has a 25 percent rate of return when advocates of competing projects claim a 100 percent rate of return?

3.4.3 Evaluation in International Organizations

International organizations could play several roles in promoting and financing rigorous evaluations. It is almost certainly counterproductive to demand that all projects be subject to impact evaluations. Clearly all projects need to be monitored to make sure that they actually happen and to avoid misuse of funds. However, some programs simply cannot be evaluated with the methods discussed in this chapter. And even among projects that could potentially be evaluated, not all need impact evaluations. In fact, the value of a poorly identified impact evaluation is very low, and its cost in terms of credibility is high, especially if international organizations take a leading role in promoting quality evaluation. A first objective is thus to cut down on the number of wasteful evaluations; any proposed impact evaluation should be reviewed by a committee before any money is spent on data collection. The committee's responsibility would be to assess the ability of the evaluation to deliver reliable causal estimates of the project's impact. A second objective would be to conduct credible evaluations in key areas. In consultation with a body of researchers and practitioners, each organization should determine key areas where it will promote impact evaluations. Randomized evaluations could also be set up in other areas when the opportunity occurs.

Credible impact evaluations require a great deal of work, and their benefits extend far beyond the organization conducting the evaluation; these factors mean that incentives to conduct rigorous evaluations are less than socially optimal. One promising remedy is to embed within the institutional framework of international agencies structures that will provide sufficient incentives for evaluators. Given the current scarcity of randomized evaluations within the institutional environment of international organizations, there may be scope for setting up a specialized unit to encourage, conduct, and finance rigorous impact

evaluations and disseminate the results. As we will briefly discuss, the poten for such a unit is tremendous: there exists a ready-made potential supply evaluators within both the international agencies and academia and collabo tions with NGOs, offering many opportunities for evaluating policies of w relevance.

Such an evaluation unit would encourage data collection and the study true natural randomized evaluations with program-induced randomizati Randomized evaluations are not the only method of conducting good imp evaluations. However, such other evaluations are conducted much more r tinely, while randomized evaluations are conducted much too rarely in the li of their value and the opportunities to conduct them. Part of the problen that no one considers conducting such evaluations to be their job, and he no one invests sufficiently to conduct them. In addition, all evaluations h common features and thus would benefit from a specialized unit with spec expertise. Since impact evaluations generate international public goods, unit should have a budget that would be used to finance and conduct rigor evaluations of internal and external projects. The unit should conduct its c evaluation projects in the key areas identified by the organization.

The unit should also work with partners, especially NGOs and academ For projects submitted from outside the unit, a committee within the unit (tentially with assisted by external reviewers) could receive proposals from w in the organization or from outsiders, and from there choose projects support. The unit could also encourage replication of important evaluati by sending out calls for specific proposals. The project could be conducted partnership with people from the unit or other researchers (academics, in p ticular). The unit could provide both financial and technical support for project, with dedicated staff and researchers. Over time, on the basis of acquired experience, the unit could serve as a more general resource center developing and diffusing training modules, tools, and guidelines (survey a testing instruments, as well as software that can be used for data entry and facilitate randomization—similar in spirit to tools produced by other units the World Bank) for randomized evaluation. The unit could also sponsor tra ing sessions for practitioners.

Another role the unit could serve, after establishing a reputation for qual is that of a dissemination agency (a clearinghouse of some sort). To be use evaluation results must be accessible to practitioners both within and outs development agencies. A key role of the unit could be to conduct systema searches for all impact evaluations, assess their reliability, and publish results in the form of policy briefs and in a readily accessible searchable da base. The database would ideally include all information that could be usefu

interpreting the results (estimates, sample size, region and time, type of project, cost, cost-benefit analysis, caveats, and so forth), as well as references to related studies. The database could include both randomized and nonrandomized impact evaluations satisfying some criteria, provided that the different types of evaluation are clearly labeled. Evaluations would need to satisfy minimum reporting requirements to be included in the database, and all projects supported by the unit would have to be included in the database, whatever their results.

Such a database would help alleviate publication bias, which may be substantial if positive results are more likely to be published. Academic journals may not be interested in publishing the results of failed programs, but from the policymakers' point of view, knowledge about negative results is just as useful as knowledge about successful projects. Comparable requirements are placed on all federally funded medical projects in the United States. Ideally, over time, the database would become a basic reference for organizations and governments, especially as they seek funding for their projects. This database could kick-start a virtuous circle, with donors demanding credible evaluations before funding or continuing projects, more evaluations being conducted, and the general quality of evaluation work rising.

3.5 Conclusion

Rigorous and systemic evaluations have the potential to leverage the impact of international organizations well beyond simply their ability to finance programs. Credible impact evaluations are international public goods: the benefits of knowing that a program works or does not work extend well beyond the organization or the country implementing the program.[7] Programs that have been shown to be successful can be adapted for use in other countries and scaled up within countries, while unsuccessful programs can be abandoned. Through promoting, encouraging, and financing rigorous evaluations (such as credible randomized evaluations) of the programs they support, as well as of programs supported by others, the international organizations can provide guidance to the international organizations themselves, as well as other donors, governments, and NGOs, in the ongoing search for successful programs. Moreover, by credibly establishing which programs work and which do not, the international agencies can counteract skepticism about the possibility of spending aid effectively and build long-term support for development. Just as randomized trials revolutionized medicine in the twentieth century, they have the potential to revolutionize social policy during the twenty-first.

Acknowledgments

This chapter draws on work that each of us has done in the field with nume coauthors, primarily in India and Kenya, respectively, and on pieces we l written synthesizing this work and discussing issues related to randomized e uations (Duflo 2004, Kremer 2003). Among other collaborators, we thank Angrist, Abhijit Banerjee, Eric Bettinger, Erik Bloom, Raghabendra Cha padhyay, Shawn Cole, Paul Glewwe, Nauman Ilias, Suraj Jacob, Elizal King, Leigh Linden, Ted Miguel, Sylvie Moulin, Robert Namunyu, Chr Vermeersch, and Eric Zitzewitz. We thank Ted Miguel for extremely deta and useful comments. We are particularly grateful to Heidi Williams for standing research assistance. We are also very grateful to François B guignon, Anne Case, Angus Deaton, Rachel Glennerster, Emily Oster, Paul Schultz.

Notes

1. There are numerous excellent technical and nontechnical surveys of these techniques as w their value and limitations. See Angrist and Krueger 1999, 2001; Card 1999; Meyer 1995.

2. Angrist and Lavy (1999) note that parents who discover they received a bad draw in the "er ment lottery" (e.g., an enrollment of thirty-eight) might then move their children out of the p school system and into private schools. However, as Angrist and Lavy discuss, private eleme schooling is rare in Israel outside the ultraorthodox community.

3. By school participation, we denote a comprehensive measure of school participation: a pu considered a participant if she or he is present in school on a given day and a nonparticipant i or he is not in school on that day.

4. Most of these papers are accessible on the IFPRI Web site www.ifpri.org.

5. One recent study not included in the analysis of Glazerman, Levy, and Meyers (2002) is th Buddlemeyer and Skoufias (2003). Buddlemeyer and Skoufias use randomized evaluation resul a benchmark to examine the performance of regression discontinuity design for evaluating the pact of the PROGRESA program on child health and school attendance and find the perform of regression discontinuity design in this case to be good.

6. Case (2001) gives an illuminating discussion of the program and the features that make its uation impossible.

7. In fact, the benefits of a credible evaluation are often negative for the person or organiz promoting the program.

References

Angrist, Joshua, Eric Bettinger, Erik Bloom, Elizabeth King, and Michael Kremer. 2 "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experime *American Economic Review* 92(5):1535–1558.

Angrist, Joshua, and Alan Krueger. 1999. "Empirical Strategies in Labor Economics." In C Ashenfelter and David Card, eds., *Handbook of Labor Economics*, Vol. 3A. Amsterdam: N Holland.

Angrist, Joshua, and Alan Krueger. 2001. "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments." *Journal of Economic Perspectives* 15(4):69–85.

Angrist, Joshua, and Victor Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *Quarterly Journal of Economics* 114(2):533–575.

Ashenfelter, Orley, Colm Harmon, and Hessel Oosterbeek. 2000. "A Review of Estimates of Schooling/Earnings Relationship, with Tests for Publication Bias." NBER Working Paper 7457. Cambridge, Mass.: National Bureau of Economic Research.

Attanasio, Orazio, Costas Meghir, and Ana Santiago. 2001. "Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to Evaluate PROGRESA." Mimeo., Inter-American Development Bank.

Banerjee, Abhijit, Shawn Cole, Esther Duflo, and Leigh Linden. 2003. "Improving the Quality of Education in India: Evidence from Three Randomized Experiments." Mimeo., Massachusetts Institute of Technology.

Banerjee, Abhijit, Suraj Jacob, and Michael Kremer, with Jenny Lanjouw and Peter Lanjouw. 2001. "Promoting School Participation in Rural Rajasthan: Results from Some Prospective Trials." Mimeo., Massachusetts Institute of Technology.

Behrman, Jere, Piyali Sengupta, and Petra Todd. 2002. "Progressing through PROGESA: An Impact Assessment of a School Subsidy Experiment in Mexico." Mimeo., University of Pennsylvania.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2002. "How Much Should We Trust Difference in Differences Estimates?" NBER Working Paper 8841. Cambridge, Mass.: National Bureau of Economic Research.

Besley, Timothy, and Anne Case. 2000. "Unnatural Experiments? Estimating the Incidence of Endogenous Policies." *Economic Journal* 110(467):F672–F694.

Bobonis, Gustavo, Edward Miguel, and Charu Sharma. 2002. "Iron Supplementation and Early Childhood Development: A Randomized Evaluation in India." Mimeo., University of California, Berkeley.

Buddlemeyer, Hielke, and Emmanuel Skofias. 2003. "An Evaluation on the Performance of Regression Discontinuity Design on PROGRESA." Discussion Paper 827. Bonn, Germany: Institute for Study of Labor.

Campbell, Donald T. 1969. "Reforms as Experiments." *American Psychologist* 24:407–429.

Card, David. 1999. "The Causal Effect of Education on Earnings." In Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics*, Vol. 3A. Amsterdam: North-Holland.

Case, Anne. 2001. "The Primacy of Education." Mimeo., Princeton University.

Cook, Thomas D. 2001. "Reappraising the Arguments Against Randomized Experiments in Education: An Analysis of the Culture of Evaluation in American Schools of Education." Mimeo., Northwestern University.

DeLong, J. Bradford, and Kevin Lang. 1992. "Are All Economic Hypotheses False?" *Journal of Political Economy* 100(6):1257–1272.

Duflo, Esther. 2001. "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment." *American Economic Review* 91(4):795–814.

Duflo, Esther. 2004. "Scaling Up and Evaluation." In Francois Bourguignon and Bais Pleskovic, eds., *Accelerating Development*. Washington, D.C. and Oxford: World Bank and Oxford University Press.

Duflo, Esther, and Emmanuel Saez. 2003. "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment." *Quarterly Journal of Economics* 118(3):815–842.

Gertler, Paul J., and Simone Boyce. 2001. "An Experiment in Incentive-Based Welfare: The Impact of PROGRESA on Health in Mexico." Mimeo., University of California, Berkeley.

Glazerman, Steven, Dan Levy, and David Meyers. 2002. "Nonexperimental Replications of Social Experiments: A Systematic Review." Interim Report/Discussion Paper. Mathematica Policy Research.

Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2003. "Teacher Incentives." NBER Wor Paper 9671. Cambridge, Mass.: National Bureau of Economic Research.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin. 2002. "Textbooks and Test Scores: Evid from a Prospective Evaluation in Kenya." Mimeo., Harvard University.

Glewwe, Paul, Michael Kremer, Sylvie Moulin, and Eric Zitzewitz. 2004. "Retrospective vs. spective Analyses of School Inputs: The Case of Flip Charts in Kenya." *Journal of Develop Economics* 74(1):251–268.

Hanushek, Eric A. 1995. "Interpreting Recent Research on Schooling in Developing Countr *World Bank Research Observer* 10:227–246.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. "Characterizing S tion Bias Using Experimental Data." *Econometrica* 66(5):1017–1098.

Heckman, James, Hidehiko Ichimura, and Petra Todd. 1997. "Matching as an Econometric E ation Estimator: Evidence from Evaluating a Job Training Program." *Review of Economic St* 64(4):605–654.

Heckman, James, Robert Lalonde, and Jeffrey Smith. 1999. "The Economics and Econometri Active Labor Market Programs." In Orley Ashenfelter and David Card, eds., *Handbook of L Economics*, Vol. 3.O. Amsterdam: North-Holland.

Heckman, James, Lance Lochner, and Christopher Taber. 1998. "General Equilibrium Treatr Effects: A Study of Tuition Policy." NBER Working Paper 6426. Cambridge, Mass.: Nati Bureau of Economic Research.

Imbens, Guido, and Joshua Angrist. 1994. "Identification and Estimation of Local Average Tı ment Effects." *Econometrica* 62(2):467–475.

Kremer, Michael. 2003. "Randomized Evaluations of Educational Programs in Developing C tries: Some Lessons." *American Economic Review Papers and Proceedings* 93(2):102–115.

Kremer, Michael, Sylvie Moulin, and Robert Namunyu. 2002. "Decentralization: a Cautio Tale." Mimeo., Harvard University.

LaLonde, Robert. 1986. "Evaluating the Econometric Evaluations of Training with Experime Data." *American Economic Review* 76(4):604–620.

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

Miguel, Edward, and Michael Kremer. 2004. "Worms: Identifying Impacts on Education Health in the Presence of Treatment Externalities." *Econometrica* 72(1):159–217.

Miguel, Edward, and Michael Kremer. 2003. "Social Networks and Learning about Healtl Kenya." Mimeo., Harvard University.

Morduch, Jonathan. 1998. "Does Microfinance Really Help the Poor? New Evidence from F ship Programs in Bangladesh." Mimeo., Princeton University.

Narayanan, Deepa (ed.) 2000. *Empowerment and Poverty Reduction: A Sourcebook*. Washing D.C.: World Bank.

Pandey, Raghaw Sharan. 2000. *Going to Scale with Education Reform: India's District Prin Education Program, 1995–99*. Washington, D.C.: World Bank.

Pitt, Mark, and Shahidur Khandker. 1998. "The Impact of Group-Based Credit Programs on F Households in Bangladesh: Does the Gender of Participants Matter?" *Journal of Political Econ* 106(5):958–996.

Pritchett, Lant. 2002. "It Pays to Be Ignorant: A Simple Political Economy of Rigorous Progı Evaluation." *Journal of Policy Reform* 5(4):251–269.

Rosenbaum, Paul R. 1995. "Observational Studies." New York: Springer.

Sen, Amartya. 2002. "The Pratichi Report." New Delhi: Pratichi India Trust.

Shultz, T. Paul. 2004. "School Subsidies for the Poor: Evaluating the Mexican PROGRESA F erty Program." *Journal of Development Economics* 74(1):199–250.

Vermeersch, Christel. 2002. "School Meals, Educational Achievement, and School Competiti Evidence from a Randomized Experiment." Mimeo., Harvard University.