Scaling Up and Evaluation

ESTHER DUFLO

This paper discusses the role that impact evaluations should play in scaling up. Credible impact evaluations are needed to ensure that the most effective programs are scaled up at the national or international levels. Scaling up is possible only if a case can be made that programs that have been successful on a small scale would work in other contexts. Therefore the very objective of scaling up implies that learning from experience is possible.

Because programs that have been shown to be successful can be replicated in other countries while unsuccessful programs can be abandoned, impact evaluations are international public goods, thus the international agencies should have a key role in promoting and financing them. In doing this, they would achieve three important objectives: improve the rates of return on the programs they support, improve the rates of return on the programs other policymakers support by providing evidence on the basis of which programs can be selected, and build long-term support for international aid and development by making it possible to credibly signal what programs work and what programs do not work.

The paper argues that considerable scope exists for expanding the use of randomized evaluations. For a broad class of development programs, randomized evaluation can be used to overcome the problems often encountered when using current evaluation practices.

Esther Duflo is the Castle Krob career development associate professor at the Massachusetts Institute of Technology in Cambridge, Mass. The author would like to thank Abhijit Banerjee, Edward Miguel, and Martin Ravallion for helpful discussions and François Bourguignon, Angus Deaton, and T. P. Schultz for extremely detailed and useful comments on a previous draft. The paper also benefited enormously from collaboration with Michael Kremer on a related paper (Duflo and Kremer forthcoming). Finally, the author is grateful to Nick Stern for asking two critical questions that greatly influenced the revision of this paper, even though this paper does not satisfactorily answer them. The author would like to thank the Alfred P. Sloan Foundation, the John D. and Catherine T. MacArthur Foundation, and the National Institutes of Health for financial support. The author is fully responsible for the content of this paper, which does not necessarily represent the views of the World Bank or of any other agency.

Annual World Bank Conference on Development Economics 2004

Introduction

Scaling up and evaluation are often presented as conflicting objectives, and for most international development agencies, "going to scale" has to be given priority. The United Nations Children's Fund (UNICEF), for example, lists as its first priority for HIV/AIDS education "moving away from small scale pilot projects" and "expanding effective and promising approaches to national scale."[1] The trade-off is explicit: by moving away from pilots and projects before their impact on behavior leading to HIV/AIDS has been convincingly established, one has to commit to expanding projects that are only promising—not enough projects have been proven to be "effective." The UNICEF web site on skill-based health education reports on 10 case studies of promising, school-based HIV/AIDS education programs, only one of which presents differences in outcomes between a treatment group and a comparison group. These approaches are the programs that UNICEF can recommend be implemented on a national scale.[2]

This paper argues that for international agencies there is no real trade-off between scaling up and evaluation. On the contrary, evaluation can give them an opportunity to leverage the impact of their programs well beyond their ability to finance them. The very idea of scaling up implies that the same programs can work in different environments and that learning from experience is possible. Therefore reliable program evaluations serve several purposes. First, a well-conducted evaluation can offer insights into a particular project. For example, all programs should be subject to process evaluations to ensure that funds are spent as intended and to receive feedback from stakeholders on how programs could be improved. However, while process evaluations are necessary, they are insufficient to determine program impact. A second purpose of rigorous evaluations of programs' impacts is that this information can be shared with others. 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) in their ongoing search for effective programs. Thus when evaluation is used to find out what works and what does not, the benefits extend far beyond the selection of projects within the organization. While a prospective impact evaluation may require postponing the national expansion of a program for some time, evaluation can be part of the backbone of a much larger expansion: that of the project on a much larger scale (if proven successful), and that of the ability to fund development projects. Providing these international public goods should be one of the important missions of international organizations.

In this paper I argue that for a broad class of development programs, randomized evaluations are a way to obtain credible and transparent estimates of program impact that overcome the problems often encountered when using other evaluation practices. Of course, not all programs can be evaluated with randomized evaluations; for example, issues such as central bank independence must rely on other methods of evaluation. Programs targeted to individuals or local communities, such as sanitation,

education, and health programs and local government reforms, are likely to be strong candidates for randomized evaluations. This paper does not recommend conducting all evaluations with randomized methods; rather, it starts from the premise that there is scope for substantially increasing their use, and that even a modest increase could have a tremendous impact.

This paper proceeds as follows: The next section presents the impact evaluation problem and the opportunities for evaluation and discusses examples of evaluations, drawn mostly from India. The following section discusses the potential of randomized evaluation as a basis for scaling up. The paper then turns to current practices and the role international agencies can play in promoting and financing rigorous evaluations, and then the paper concludes.

The Methodology of Randomized Evaluation

This section discusses the methodology of randomized evaluation: the problem it tries to solve, and the solution it provides. It presents various examples where the methodology is applied.

The Evaluation Problem

Any impact evaluation attempts to answer essentially counterfactual questions: How would individuals who did benefit from the program have fared in the absence of the program? How would those who did not benefit have fared if they had been exposed to the program? The difficulty with these questions is immediate: at a given point in time, an individual is observed either exposed to the program or not exposed. Comparing the same individual over time will not, in most cases, provide a reliable estimate of the impact the program had on him or her, because many other things may have changed at the same time that 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 with a similar group that was not exposed to the program.

Thus the critical objective of impact evaluation is to establish a credible comparison group, that is, 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 gives us an idea of what would have happened to the program group if it had not been exposed, and thus allows us to obtain an estimate of the average impact on the group in question. Generally, in the real world, individuals who were subjected to the program and those who were not are very different, because programs are placed in specific areas (for example, poorer or richer areas), individuals are screened for participation in the program (for instance, on the basis of poverty or on the basis of their motivation), and the decision to participate is often voluntary. For all these reasons, those who were not exposed to a program are often not comparable to those who were. Any difference between them could be attributed to two factors: preexisting

differences (the so-called selection bias) and the impact of the program. Because we have no reliable way to estimate the size of the selection bias, we 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 to determine which group is a likely comparison group. One situation where the selection bias disappears is when the treatment and the comparison groups are selected randomly from a potential population of beneficiaries (individuals, communities, schools, or classrooms can be selected into the program). In this case, on average, we can be assured that those who are exposed to the program are no different than those who are not, and that a statistically significant difference between them in the outcomes that the program was planning to affect after the program is in place can be confidently attributed to the program. I will now discuss several examples of randomized evaluations.

Prospective Randomized Evaluations

Random selection of treatment and comparison groups can happen in several circumstances: during a pilot project because the program's resources are limited, or because the program itself calls for random beneficiaries. The next two subsections discuss examples of these different scenarios. In addition, in some circumstances where a program was not randomly allocated, because of favorable circumstances a credible control group nevertheless exists.

Pilot Projects

Before a program is launched on a large scale, a pilot project, necessarily limited in scope, is often implemented. In most circumstances, the beneficiaries of the pilot can be randomly chosen, because many potential sites (or individuals) are equally good locations for the project. The pilot can then be used not only to find out if the program turns out to be feasible (which is what most pilots are currently used for) but also whether the program has the expected impacts. Job training and income maintenance programs are prominent examples of randomized evaluations. A growing number of such pilot projects are evaluated, often by means of collaboration between an NGO and academics (see, for example, Kremer 2003 for several references). To illustrate briefly how such studies can work in practice, consider an example from India analyzed in Banerjee and others (2001). This study evaluated a program where an Indian NGO, Seva Mandir, decided to hire second teachers for the nonformal education centers it runs in villages. 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. These centers are plagued by high teacher and child absenteeism. A second teacher, often a woman, was randomly assigned to 21 out of 42 schools. The hope was to increase the number of days the school was open, to increase children's participation, and to improve 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 and child attendance were regularly monitored in program and comparison

schools for the entire duration of the project. The impact of the program on learning was measured by testing children at the end of the school year. The program reduced the number of days schools were closed: one-teacher schools were closed 39 percent of the time, whereas two-teacher schools were closed 24 percent of the time. Girls' attendance increased by 50 percent. However, test scores did not differ.

Carefully evaluated pilot projects form a sound basis for the decision to scale the project up. In the example just discussed, the NGO did not implement the two-teacher program on a full scale on the grounds that its benefits did not outweigh its costs. The NGO used the savings to expand other programs.

By contrast, positive results can help build consensus for a project that has the potential to be extended far beyond the scale that was initially envisioned. The PROGRESA program (subsequently renamed as PROGRESA-Oportunidades) in Mexico is the most striking example of this phenomenon.[3] PROGRESA offers grants, distributed to women, conditional on children's school attendance and preventative health measures (nutrition supplementation, health care visits, and participation in health education programs). In 1998, when the program was launched, Mexican government officials made a conscious decision to take advantage of the fact that budgetary constraints made it impossible to reach the 50,000 potential beneficiary communities of PROGRESA all at once, and instead started with a pilot program in 506 communities. Half of those 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 starting with this pilot program was to increase the probability that the program would be continued in case of a change in the party in power. The proponents of the program understood that to be scaled up successfully, the program would require continuous political support.

The task of evaluating the program was given to academic researchers through the International Food Policy Research Institute. The data were made accessible to many different people, and a number of papers have been written on the program's impact (most of them are accessible on the institute's web site).[4] The evaluations showed that PROGRESA was effective in improving health and education. Comparing PROGRESA beneficiaries and nonbeneficiaries, Gertler and Boyce (2001) show that children had about a 23 percent reduction in the incidence of illness, a 1 to4 percent increase in height, and an 18 percent reduction in anemia. Adults experienced a reduction of 19 percent in the number of days of work lost because of illness. Shultz (forthcoming) finds an average of a 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: 14.8 percent.

In part because the program had been shown to be successful, it was indeed maintained when the Mexican government changed hands: by 2000 it 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 Mexico's gross domestic product (Gertler and Boyce 2001). It was subsequently expanded to urban communities, and, with support from the World Bank, several neighboring Latin American countries are implementing similar programs. Mexican officials transformed a budgetary constraint into an opportunity and made evaluation the cornerstone of subsequent scaling up. They were

rewarded by the expansion of the program and by the tremendous visibility that it acquired.

Replication and Evaluation of Existing Projects

A criticism often heard against the evaluation of pilot projects is that pilot projects may be different from regular projects. The fact that pilot projects are evaluated can create problems with the interpretation of the results. If the project is unsuccessful, it may be because it faced implementation problems during the first phase of the program. If it is successful, it may be because more resources were allocated to it than would have been under a more realistic scenario, because the context was favorable, or because the participants in the experiment had a sense of being part of something and changed their behavior. Moreover, all programs are implemented under particular circumstances, and the conclusions may be hard to generalize.

A first answer to some of these concerns is to replicate successful (as well as potentially successful) experiments in different contexts. This has two advantages. First, in the process of transplanting a program, circumstances will require changes, and the program will show its robustness if its effectiveness survives these changes. Second, obtaining several estimates in different contexts will provide some guidance about whether the impacts of the program differ significantly for different groups. Replication of the initial evaluation study in the new context does not imply delaying full-scale implementation of the program if the latter is justified on the basis of existing knowledge. More often than not, the introduction of the program can only proceed in stages, and the evaluation only requires that beneficiaries be phased into the program in random order.

Two studies on school-based health interventions provide a good illustration of these two benefits. The first study (Miguel and Kremer 2004) evaluates a program of twice-yearly, school-based mass treatment with inexpensive deworming drugs in Kenya, where the prevalence of intestinal worms among children is 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 because of reduced disease transmission. Absenteeism in treatment schools was 25 percent (or 7 percentage points) lower than in comparison schools. Including this spillover effect, the program increased schooling by 0.15 years per person treated. Combined with estimates about the rates of return to schooling, the estimates suggest extremely high rates of return of the deworming intervention. The authors estimate that deworming increases the net present value of wages by more than US$30 per treated child at a cost of only US$0.49.

One of the authors then decided to examine whether these results generalized among preschoolers in urban India (Bobonis, Miguel, and Sharma 2002). The baseline revealed that, even though worm infection was present, the levels of infection were substantially lower than in Kenya (in India, "only" 27 percent of children suffer from some form of worm infection, compared with 92 percent in Kenya). However, 70 percent of children had moderate to severe anemia. The program was therefore modified to include iron supplementation. It was administered through a network of preschools

in urban India. After one year of treatment, the authors found a nearly 50 percent reduction in moderate to severe anemia, large weight gains, and a 7 percent reduction in absenteeism among four- to six-year-olds (but not for younger children). The results of the previous evaluation were thus by and large vindicated.[5]

A second answer is to evaluate the impact of programs that have already shown their potential to be implemented on a large scale. In this case, concerns about the ability to expand the program are moot, at least at the level at which it was implemented. It may also make evaluating the program in several sites at the same time easier, thereby alleviating some of the concerns about external validity. A natural occasion for such evaluation is when the program is ready to expand, and the expansion can be phased in in random order.

The evaluation of a remedial education program by Banerjee and others (2003) is an example of this approach. The program has been run by Pratham, an Indian NGO, which implemented it in 1994. Pratham now reaches more than 161,000 children in 20 cities. The remedial education program hires a young woman from the children's community to provide remedial education in government schools to children who have reached grades 2, 3, or 4 without having mastered the basic grade 1 competencies. Children who are identified as lagging behind are pulled out of their regular classroom for two hours a day to receive this instruction. Pratham wanted to evaluate the impact of this program, one of their flagship interventions, at the same time as they were looking to expand. The expansion into a new city, Vadodara, provided an opportunity to conduct a randomized evaluation. In the first year (1999–2000), the program was expanded to 49 randomly selected schools out of the 123 Vadodara government schools. In 2000–01, the program was expanded to all the schools, but half the schools got a remedial teacher for grade 3 and half got one for grade 4. Grade 3 students in schools that got the program in grade 4 served as the comparison group for grade 3 students in schools who got the program in grade 3. At the same time, a similar intervention was conducted in a district of Mumbai, where half the schools got the remedial teachers in grade 2 and half got them in grade 3. The program was continued for one more year, with the schools switching groups. Thus the program was conducted in several grades, in two cities, and with no school feeling that it had been deprived of resources relative to the others, because all schools benefited from the program.

After two years, the program increased the average test score by 0.39 standard deviations, which represents an increase of 3.2 points out of a possible 100 (the mean in the control group was 32.4 points), and had an even stronger impact on the test scores of those children who had low scores initially (an increase of 3.7 points, or 0.6 standard deviation, on a basis of 10.8 points). The impact of the program is rising over time, but it is similar across cities and genders. Hiring remedial education teachers from the community appears to be 10 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, because the program was continued for a period of time, it was evaluated in two quite different contexts, and it has shown its ability to be rolled out on a large scale.

Program-Induced Randomization

In some instances, fairness or transparency considerations make randomization the best way to choose the recipients of a program. Such programs are natural candidates for evaluation, because the evaluation exercise does not require any modification of the program's design.

When some schools are oversubscribed, allocation to particular schools is often done by lottery. In some school systems in the United States, students have the option of applying to so-called magnet schools or schools with special programs, and admission is often granted by lottery. Cullen, Jacob, and Levitt (2002) use this feature to evaluate the impact of school choice in the Chicago school system by comparing lottery winners and losers. Because each school runs its own lottery, their paper is, in effect, taking advantage of 1,000 different lotteries. They find that lottery winners are less likely to attend their neighborhood schools than lottery losers, but more likely to remain in the Chicago school system. However, their subsequent performance is actually worse than that of lottery losers. This is in sharp contrast to expectations and what a "naive" comparison would have found. When one simply compares the results of all the children who attended the school of their choice to the results of all those who did not, one finds that the results of children who attended a school of their choice are indeed better than the results of those who did not. The results from the randomized evaluation show, however, that, if anything, the causal effect of attending a school of one's choice is negative. The "naive" difference, which is positive, simply reflects the fact that the children who decided to change schools were highly motivated.

Voucher programs constitute another example of programs that often feature a lottery. The sponsor of the program allocates only a limited budget to the program, the program is oversubscribed, and a lottery is used to pick the beneficiaries. Angrist and others (2002) evaluate a Colombian program in which vouchers for private schools were allocated by lottery because of the program's limited budget. Vouchers were renewable conditional on satisfactory academic performance. The authors compare lottery winners and losers. Lottery winners were 15 to 20 percent more likely to attend private school; 10 percent more likely to complete grade 8; and scored 0.2 standard deviations higher on standardized tests, equivalent to a full grade level. Winners were substantially more likely to graduate from high school and scored higher on high school completion and college entrance examinations. The benefits of this program to participants clearly exceeded the costs, which were similar to the costs of providing a public school place.

When nationwide policies include some randomization aspect, this provides a unique opportunity to evaluate a policy that has already been scaled up in several locations. The knowledge gained from this experience can be used to inform policy decisions to expand the policy in the countries, to continue with the program, or to expand the policy in other countries. However, because the randomization is part of the program design rather than a deliberate attempt to make evaluating it possible, the data necessary for the evaluation are not always available. International agencies

can play two key roles in this respect. First, they can organize and finance limited data collection efforts. Second, they can encourage governments and statistical offices to link up existing data sources that can be used to evaluate the experiments. Set-asides for women and minorities in the decentralized government in India (the *panchayat* system) are an interesting example. In 1993, the 73rd amendment to the Constitution of India required the states to set up a three-tiered *panchayat* system (village, block, and district levels), directly elected by the people, to administer local public goods. Elections must take place every five years, and *panchayat* councils have the latitude to decide how to allocate local infrastructure expenditures. The amendment also required that one-third of all positions (council members and council chairs) be reserved for women, and that a share equal to the representation of disadvantaged minorities (scheduled castes and scheduled tribes) be reserved for these minorities. To avoid any possible manipulation, the law stipulated that the reserved positions be randomly allocated.

Chattopadhyay and Duflo (forthcoming) evaluate the impact of reserving seats for women in West Bengal. They collected data on 465 villages across 165 councils in 1 district, and find that women tend to allocate more resources to drinking water and roads and less to education. This corresponds to the priorities men and women expressed through their complaints to *panchayats*. Then they collected the same data in a poor district of Rajasthan, Udaipur. They find that in Rajasthan, women invest more in drinking water and less on roads, and that this once again corresponds to the complaints expressed by men and women. These results were obtained in two very different districts with different histories: West Bengal had had a *panchayat* since 1978, while Rajasthan had none until 1995, plus Rajasthan has particularly low female literacy among Indian states. Thus the results suggest that the gender of policymakers matters in both more and less developed political systems. Furthermore, it provides indirect, but powerful, evidence that local elected officials do have power even in relatively young systems. Chattopadhyay and Duflo (forthcoming) also evaluate the impact of reservations for scheduled castes and find that a larger share of goods is assigned to scheduled caste hamlets when the head of a *panchayat* is from a scheduled caste.

In principle, the data to evaluate the impact of this experiment on a much larger scale are available: village-level census data are available for 1991 and will become available for 2001. The National Sample Survey Organization conducts large-scale consumption and labor surveys every five years, with detailed data on outcomes. However, administrative barriers make these data difficult to use for the purpose of evaluating this program, because the census does not contain any information about which *panchayat* a village belongs to. In addition, the information about *panchayat* reservations and composition is not centralized, even at the state level, and is available only at the district level. Likewise, the National Sample Survey contains no information about *panchayats*. This is an example where, at a relatively small cost, information could be made available that would be useful for evaluating an extremely large program. It requires coordinating various people and agencies, a task that international organizations should be well placed to accomplish.

Other Methods to Control for Selection Biases

Natural or organized randomized experiments are not the only methodology that can be used to obtain credible impact evaluation of program effects. To compensate for the lack of randomized evaluations, researchers have developed alternative techniques to control for selection bias as best as possible. Labor economists in particular have made tremendous progress. (For excellent technical and nontechnical surveys of the various techniques, their value, and their limitations, see, for example, Angrist and Krueger 1999, 2001; Card 1999; and Meyer 1995.) Here I briefly mention some of the techniques that are most popular with researchers.

One strategy is to try to find a control group that is as comparable as possible with the treatment group, at least along observable dimensions. This can be done by collecting as many covariates as possible, and adjusting the computed differences through a regression or by matching the program and the comparison group, that is, by forming a comparison group that is as similar as possible to the program group. One way to proceed is to predict the probability that a given individual is in the comparison or the treatment group on the basis of all the available observable characteristics and to form a comparison group by picking people who have the same probability of being treated as those who actually got treated. Rosenbaum (1995) refers to this as 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 treatment and control groups. In cases where treatment is assigned on the basis of a variable that is not observed by the researcher, such as demand for the service, this technique will lead to misleading inferences.

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 (this is often called the difference-in-differences technique). Whether the argument is good and the identification assumptions are justified is, however, often hard to judge. This identification assumption cannot be tested, and to even ascertain its plausibility, one needs to have long time series of data from before the program was implemented 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 needs to take into account that regions are often affected by time-persistent shocks, which may look like a program effect (Bertrand, Duflo, and Mullainathan 2004).

Duflo (2001) takes advantage of a rapid school expansion program that took place in Indonesia in the 1970s to estimate the impact of building schools on schooling and subsequent wages. Identification is made possible because the allocation rule for schools is known (more schools were built in places with low initial enrollment rates), and because the cohorts benefiting from the program are easily identified (children 12 or older when the program started did not benefit from the program). The faster growth of education across cohorts in regions that got more schools suggests that access to schools contributed to increased education. The trends were similar before the program and shifted clearly for the first cohort that was exposed to the program,

which reinforces confidence in the identification assumption. This identification strategy is not often valid, however. Frequently when policy changes are used to identify the effect of a particular policy, the policy change is itself endogenous to the outcomes it tried to affect, which makes identification impossible (see Besley and Case 2000).

Program rules often generate discontinuities that can be used to identify the effects of the program by comparing those who were just above the threshold to qualify for a program to those who were just below the threshold. For example, if scholarships are allocated on the basis of a certain number of points, one can compare those just above to those just below the threshold. Angrist and Lavy (1999) use this technique, known as regression discontinuity design (Campbell 1969) to evaluate the impact of class size in Israel. In Israel, a second teacher is allocated whenever the size of a class would be larger than 40 children. This generates discontinuities in class size when the enrollment in a grade goes from 40 to 41 (class size changes from 40 to 20 and 21), 80 to 81, and so on. Angrist and Lavy compare test score performance in schools just above and just below the threshold and find that those just above the threshold have significantly higher test scores than those just below, which can confidently be attributed to the class size, because it is difficult to imagine that schools on both sides of the threshold have any other systematic differences. Discontinuities in program rules, when enforced, are thus a source of identification. However, they often are not enforced, especially in developing countries. For example, researchers tried to use the discontinuity in Grameen Bank, the flagship micro-credit organization in Bangladesh that lends only to people who own less than one acre of land (Pitt and Khandker 1998), as a source of identification. However, in practice Grameen Bank lends to many people who own more than one acre of land, and there is no discontinuity in the probability for borrowing at the threshold (Morduch 1998). In developing countries rules are probably frequently not enforced strictly enough to generate discontinuities that can be used for identification purposes.

Alternatives to randomized evaluation exist, and they are useful; however, identification issues need to be tackled with extreme care and they are never self-evident. They generate intense debate in academic circles whenever such a study is conducted. Identification is less transparent, and more subject to divergence of opinion, than in the case of randomized experiments. The difference between good and bad evaluations of this type is thus more difficult to communicate. The study and the results are also less easy to convey to policymakers in an effective way with all the caveats that need to accompany them. This suggests that, while a mix of randomized and nonrandomized evaluation is necessary, international organizations should commit themselves to running some randomized evaluations.

Scaling Up and Randomized Evaluations

The previous section showed that when programs' beneficiaries are individuals or communities, rather than an entire country, for example, randomized evaluations are

often a possible way to obtain reliable estimates of the programs' effects. This section discusses how the results of these evaluations can be used to scale up development programs.

Obtaining Reliable Estimates of Program Impact

When the evaluation is not planned ex ante, to evaluate the impact of a program researchers must resort to before and after comparisons (when a baseline was conducted), or to comparisons between beneficiaries and communities that, for some reason, were not exposed to the program. When the reasons why some people were exposed to the program and some were not are unknown, or worse, when they are known to be likely to introduce selection bias, those comparisons are likely to be biased. The data collection is often as expansive as for a randomized evaluation, but the inferences are biased. As argued earlier, controlling for observable differences between treatment and control groups through a regression analysis or propensity score matching will correct for the bias only if beneficiaries and nonbeneficiaries are known with certainty to be comparable conditional on these characteristics. This is unlikely to be true unless the program was randomly allocated conditional on these characteristics. In particular, a project officer trying to optimally allocate a program typically has more information than a researcher, and will (and should) make use of it when allocating resources.

These concerns have serious practical implications. Studies comparing experimental and nonexperimental estimates with the same data show that the results from randomized evaluation can be quite different from those drawn from nonrandomized evaluation. In a celebrated analysis of job training programs, LaLonde (1986) finds 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 differed significantly from experimental results. Glewwe and others (forthcoming) compare retrospective and prospective analyses of the effect of flip charts in schools on test scores. Retrospective estimates using straightforward ordinary least squares regressions suggested that flip charts raised test scores by up to 20 percent of a standard deviation, robust to the inclusion of control variables, while difference-in-differences estimates suggested a smaller effect 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 estimates based on randomized evaluations provided no evidence that flip charts increased test scores. These results suggest that using retrospective data to compare test scores seriously overestimates the charts' effectiveness. A difference-in-differences approach reduced, but did not eliminate, the problem and, moreover, whether such a difference-in-differences approach has general applicability is not clear. These examples suggest that ordinary least squares estimates are biased upward rather than downward. This is plausible, because 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 biases were similar in other retrospective analyses of education inputs in developing countries, the effects of inputs may be even more modest than retrospective studies suggest.

Some of the results are more encouraging. For example, Buddelmeyer and Skoufias (2003) use randomized evaluation results as a benchmark to examine the performance of regression discontinuity design for evaluating the impact of the PROGRESA program on child health and school attendance. The researchers found the performance of regression discontinuity design in this case to be remarkably good: impact estimates with this quasi-experimental method agreed with experimental evidence in 10 out of 12 cases, and the 2 exceptions both occurred in the first year of the program. Such research can provide invaluable guidance about the validity and potential biases of quasi-experimental estimators.

Another important source of bias in program effects are publication biases. Positive results tend to receive a large amount of publicity. Agencies that implement programs seek publicity for their successful projects, and academics, as well as academic journals, are much more interested in and able to publish positive results than modest or insignificant results. However, many programs fail, and publication bias may be substantial if only positive and significant results are published.

The problem of publication bias may be much larger with retrospective evaluations. Ex post the researchers or evaluators define their own comparison group, and thus may be able to pick a variety of plausible comparison groups. In particular, researchers obtaining negative results with retrospective techniques are likely to try different approaches or not to publish.

Available evidence suggests that the publication bias problem is severe (DeLong and Lang 1992). In the case of natural experiments and instrumental variable estimates, publication bias may actually more than compensate for the reduction in bias caused by the use of an instrument, because they tend to have larger standard errors, and researchers looking for significant results will select only large estimates. For example, Ashenfelter, Harmon, and Oosterbeek (1999) provide strong evidence of publication bias of instrumental variables 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 results that claim 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 to conduct a prospective randomized evaluation has been done, researchers just need to ensure that the results are documented and published even if the results suggest quite modest effects, or even no effects at all (such as some of the studies discussed in this paper). As I discuss later, putting institutions in place to ensure that negative results are systematically disseminated is important (such a system is already in place for the results of medical trials).

Several sources of bias are specific to randomized evaluation, but they are well known and can often be corrected for. The first possibility is that the initial randomization is not respected; for example, a local authority figure insists that the school in his village be included in the group scheduled to receive the program, or parents

manage to reallocate their children from a class or a school without the program to a class or school with the program. Or conversely, individuals allocated to the treatment group may not receive the treatment, for example, because they decide not to take the program up. Even though the intended allocation of the program was random, the actual allocation is not. In particular, the program will appear to be more effective than it is in reality if individuals allocated to the program ex post also receive more of other types of resources, which is plausible. This concern is real, and evaluations certainly need to deal with it; however, it can be dealt with relatively easily. Even though the initial assignment does not guarantee in this case that someone is actually either in the program or in 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 (this difference is often called the intention to treat estimate) and scale up the difference by dividing it by the difference in the probability of receiving the treatment in those two groups (Imbens and Angrist 1994). Krueger's (1999) reanalysis of the Tennessee student/teacher achievement ratio class size experiment used this method to deal with the fact that some parents had managed to reallocate their children from regular classes to small classes.[6] Such methods will provide an estimate of the average effect of the treatment on those who were induced to take the treatment by the randomization, for instance, on children who would have been in a large class had they not been placed in the treatment groups. This may be different from the average effect in the population, because people who anticipated benefiting more from the program may be more likely to take advantage of it. It may, however, be a group that policymakers especially care about, because they are likely to be the ones who are more likely to take advantage of the policy if it is implemented on a large scale.

A second possible source of bias is differential attrition in the treatment and comparison groups: those who benefit from 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 Banerjee and others (2001) analyze 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 by the program from dropping out of school are the weakest in the class, the comparison between the test scores of the children in treatment and control schools may be biased downward. Statistical techniques can be used to deal with this problem, but the most effective way is to try to limit attrition as much as possible. For example, in the evaluation of the remedial education program in India (Banerjee and others 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 villages were not tested. As a result, the attrition rate remained relatively high, but was the same in the treatment and comparison schools and does not invalidate test score comparisons.

A third possible source of bias is when the comparison group is itself indirectly affected by the treatment. For example, Miguel and Kremer's (2004) study of the Kenyan deworming program showed that children in treatment schools and in

schools near the treatment schools were less likely to have worms, even if they were not themselves given the medicine. The reason is that worms easily spread from one person to another. In previous evaluations, treatment had been randomized within schools. Its impact was thus underestimated, because even comparison children benefited from the treatment. The solution in this case was to choose the school rather than the pupils within a school as the unit of randomization.

Randomizing across units—for example, across schools or communities rather than individuals within a unit—is also often the only practical way to proceed. For example, offering a program to some villagers and not others may be impossible, but the fact that randomization takes places at the group rather than the individual level needs to be explicitly taken into account when calculating the confidence interval of the estimates of the impact of the program. Imagine, for example, that only two large schools take part in a study, and that one school is chosen at random to receive new textbooks. The differences in test scores between children in the two schools may reflect many other characteristics of the treatment and comparison schools (for example, the quality of the principal). Even if the sample of children is large, the sample of schools is actually small. The grouped nature of the data can easily be taken into account, but it is important to take it into account when planning design and sample size.

In summary, while randomized evaluations are not a bullet-proof strategy, the potential for biases is well known, and those biases can often be corrected. This stands in sharp contrast with biases of most other types of studies, where the bias caused by the nonrandom treatment assignment cannot be either signed or estimated.

Generalizing the Results of Evaluation

Randomized evaluation can therefore provide reliable estimates of treatment effects for the program and the population under study. To draw on these estimates to assess the prospects for scaling up the program, however, one has to make the case that these estimates tell us something about the effect of the program after it is scaled up. There are different reasons why the results of a well-executed experiment may not be generalizable.

First, the experiment itself may have affected the treatment or the comparison samples, for example, the provision of inputs might temporarily increase morale among beneficiaries, and this could improve performance (known as the Hawthorne effect). While this would bias randomized evaluations, it would also bias fixed-effect or difference-in-differences estimates. As mentioned earlier, either the treatment or the comparison group may also be temporarily affected by being part of an experiment (known as the John Henry effect). These effects are less likely to be present when the evaluations are conducted on a large scale and over a long enough time span, and some experimental designs can minimize the risk of such effects. For example, in Pratham's remedial education program analyzed by Banerjee and others (2003), all the schools received the program, but not all the grades. Trying to assess whether these effects are present is, however, important. In his reanalysis of

the project student/teacher achievement ratio data, Krueger (1999) exploits variation in class size within the control group occasioned by children's departure during the year to obtain a second estimate of the class size effect, which is, by definition, not contaminated by John Henry or Hawthorne effects, because all the teachers in this sample belong to the control group. He finds no difference in the estimates obtained by these two methods.

Second, treatment effects may be affected by the scale of the program. For example, the Colombian voucher program Angrist and others (2002) analyze was implemented on a pilot basis with a small sample, but the rest of the school system remained unchanged, in particular, the number of students affected was too small to have an impact on the composition of the public and private schools. If this program were to be implemented on a large scale, it could affect the functioning of the school system, and could therefore have a different impact (Hsieh and Urquiola 2002). More generally, partial equilibrium treatment effects may be different from general equilibrium treatment effects (Heckman, Lochner, and Taber 1998). Addressing these problems requires randomized evaluation to be performed at the level of the economy. This may be possible for programs such as voucher programs, where the general equilibrium effects will likely take place at the community level, and where communities can be randomly affected or not affected by the program, but I am not aware of an evaluation of this type.

Third, and perhaps most important, no project will be replicated exactly: circumstances vary and any idea will have to be adapted to local circumstances. In other words, internal validity is not sufficient. The evaluation also needs to have some external validity, that is, the results can be generalized beyond the population directly under study. Some argue that evaluation can never generalize. In its most extreme form (see, for example, Cronbach 1982; Cronbach and others 1980; see also the review of the education literature in Cook 2001), this argument contends that every school, for example, is specific and complex, and that nothing definitive can be learned about schools in general. This discourse has made its way into some international organizations,[7] but note that it is contradictory to the objective of going to scale. What is the point of rolling out a program on a large scale if one thinks that, for example, each school needs a different program? The very objective of scaling up has to be founded on the postulate that even if the impact of a program varies across individuals, thinking of average treatment effects makes sense. This is exactly the postulate that underlies the external validity of randomized evaluations.

A theory of why a specific program is likely to be effective is necessary to provide some guidance about what elements in the program and in its context were keys to its success. Theory will help disentangle the distinct components of a program and discriminate between variants that are likely to be important and variants that are not (Banerjee 2002). For example, an economic analysis of the PROGRESA program suggests that it may have been useful because of its impact on income, because of its effect on women's bargaining power, or because of its effect on incentives. Aspects of the program most likely to be relevant to the program's success are the size of the transfer, its recipient, and the conditionality attached to it. In contrast, the color of

the food supplement distributed to the families, for example, is unlikely to be important. Replication of the programs may then vary these different aspects to determine which of them is the most important. This also suggests that priority should be given to evaluating programs that are justified by some well-founded theoretical reasoning, because the conclusions from the evaluation are then more likely to generalize.

Theory provides some guidance about what programs are likely to work and, in turn, the evaluation of these programs forms a test of the theory's prediction. Because prospective evaluations need to be planned ahead of time, designing pilot programs in such a way that they help answer a specific question or test a specific theory is also often possible. For example, Duflo, Kremer, and Robinson (2003) report on a series of randomized evaluations conducted in Kenya in collaboration with International Christian Support (ICS) Africa, a Netherlands-based NGO active in the area. They were motivated by the general question: why do so few farmers in this region of Kenya use fertilizer (only about 10 percent), even though its use seems to be profitable and it is widely used in other developing countries, as well as in other regions of Kenya? They first conducted a series of trials on the farms owned by randomly selected farmers and confirmed that, in small quantities, fertilizer is extremely profitable: the rates of return were often in excess of 100 percent. They then conducted a series of programs to answer a number of other questions: Do farmers learn when they try fertilizer out for themselves? Do they need information about returns or about how to use them? Does the experiment need to take place on their farm, or can it take place on a neighbor's farm? Do they learn from their friends? To answer these questions, the researchers first randomly selected farmers to participate in the field trials and followed their adoption of fertilizer subsequently, as well as that of a comparison group. Second, they also followed adoption by the friends and neighbors of the comparison farmers. Finally, they invited randomly selected friends of farmers participating in the trials to the important stages in the development of the experiment and monitored their subsequent adoption.

These questions are extremely important to our understanding of technology adoption and diffusion, and the ability to generate exogenous variation through randomized program evaluation greatly helped in this understanding. Moreover, the answers also helped International Christian Support Funds develop a school-based agricultural extension program that has a chance to be effective and cost-effective. A pilot version of this program is currently being evaluated.

Thus theory and existing evidence can be used to design informative replication experiments and to sharpen predictions from these experiments. Rejection of these predictions should then be taken seriously and will inform the development of the theory. Replication is one area where international organizations, which are present in most countries, can play a key role if they take the time to implement randomized evaluations of programs that can be replicated. An example of such an opportunity that was seized is the replication 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 a randomized evaluation and are currently being evaluated.

Note also that the exogenous variation created by the randomization can be used to help identify a structural model. Attanasio, Meghir, and Santiago (2001) and Behrman, Sengupta and Todd (2002) are two examples of such an exercise using the PROGRESA data to predict the possible effects of varying the schedule of transfers. These studies rest on assumptions that one is free to believe or not, but at least they are freed of some assumption by the presence of this exogenous variation. The more general point is that randomized evaluations do not preclude the use of theory or assumptions. Indeed, they generate data and variation that can help identify some aspects of these theories.

Assessing the Feasibility of Randomized Evaluation

As noted in the introduction, randomized evaluations are not adapted for all types of programs. They are adapted to programs that are targeted to individuals or communities and where the objectives are well defined. For example, the efficacy of foreign aid disbursed as general budget support cannot be evaluated in this way. It may be desirable, for efficiency or political reasons, to disburse some fraction of aid in this form, although it would be extremely costly to distribute all foreign aid in the form of general budget support, precisely because it leaves no place for rigorous evaluation of projects. However, in many cases randomized evaluations are feasible.

The main cost of evaluation is the cost of data collection, and it is no more expensive than the cost of collecting any other data. Indeed, by imposing some discipline on which data to collect (the outcomes of interest are defined ex ante and do not evolve, as the program fails to affect them) may reduce the cost of data collection relative to a situation where what is being measured is not clear. Several potential interventions can also be evaluated in, say, the same groups of schools, as long as the comparison and treatment groups for each intervention are "criss-crossed." This has the added advantage of making is possible to directly compare the efficacy of different treatments. For example, in Vadodara, Pratham implemented a computer-assisted learning program in the same schools where the remedial education program evaluated by Banerjee and others (2003) was implemented. The program was implemented only in grade 4. Half the schools that had the remedial education program in grade 4 got the computer-assisted learning program, and half the schools that did not have the remedial education program got the computer-assisted learning program. The preliminary results suggest that the effect on mathematics is comparable to the effect of the remedial education program, but the cost is much smaller. Even keeping the budget of process evaluation constant, a reallocation of part of the money that is currently spent on unconvincing evaluation would probably go a long way toward financing the same number of randomized evaluations. Even if randomized evaluations turn out to be more expensive, the cost is likely to be trivial in comparison with the amount of money saved by avoiding the expansion of ineffective programs. This suggests that randomized evaluation should be financed by international organizations.

Political economy concerns sometimes make not implementing a program in the entire population difficult, especially when its success has already been demonstrated;

for example, the urban version of PROGRESA will not start with a randomized evaluation, because of the strong opposition to delaying some people's access to it. This objection can be tackled at several levels. First, opposition to randomization is less likely to falter in an environment where it has strong support, especially if a rule prescribes that an evaluation is necessary before full-scale implementation.

Second, if, as argued earlier, evaluations are not financed by loans but by grants, this may make it easier to convince partners of their usefulness, especially if they permit countries to expand programs. An example of such explicit partnership is a study on the effectiveness of HIV/AIDS education currently being conducted in Kenya (Duflo and others 2003). With support from UNICEF, the government of Kenya has put together a teacher training program for HIV/AIDS education. Because of a lack of funds, the program's coverage had remained minimal. The Partnership for Child Development, with grants from the World Bank, is funding a randomized evaluation of the teacher training program. ICS Africa is organizing training sessions with facilitators from the Kenyan government. The evaluation allowed training to be expanded to 540 teachers in 160 schools, which would not have been possible otherwise. The randomization was not grounds for the Kenyan authorities to reject the program. On the contrary, at a conference organized to launch the program, Kenyan officials explicitly appreciated the opportunitythe evaluation gave them to be at the forefront of efforts to advance knowledge in this area.

The example of PROGRESA showed that government officials recognized the value of randomized evaluation and were actually prepared to pay for it. The favorable response to PROGRESA and the World Bank's subsequent endorsement of the findings will certainly influence how other governments think about experiments. Several examples of this kind could do a lot to move the culture.

Third, governments are far from being the only possible outlets through which international organizations could organize and finance randomized evaluation. Many of the evaluations discussed so far were set up in collaboration between NGOs and academics. NGOs have limited resources and therefore cannot hope to reach all the people they target. Randomized allocation is often perceived as a fair way to allocate sparse resources. In addition, members of NGOs are often extremely entrepreneurial, and as a result NGOs are willing to evolve in response to new information. NGOs tend to welcome information about the effectiveness of their programs, even if they find out that they are ineffective. For these reasons, many NGOs are willing to participate in randomized evaluations of their programs. For example, the collaboration between the Indian NGO Pratham and Massachusetts Institute of Technology researchers, which led to the evaluations of the remedial education and the computer-assisted learning program (Banerjee and others 2003) was initiated by Pratham, which was looking for partners to evaluate their program. Pratham understood the value of randomization and was able to convey it to the schoolteachers involved in the project. International organizations could finance randomized evaluations organized in collaboration with researchers (from their organizations or from academia) and genuine NGOs.

Timing Evaluation and Implementation

Prospective evaluations do take time: convincing studies often go on for two or three years. Obtaining information about a program's long-term impact, which can be extremely important and can differ from the short-run impact, takes even longer. For example, Glewwe, Illias, and Kremer (2003) suggest that a teacher incentive program caused a short-run increase in test scores but no long-run impact, which they attribute to practices of "teaching to the test." When the program targets children but seeks to affect adult outcomes, which is the case for most education or health interventions, the delay between the program and the outcomes may become long. In these cases, waiting for the answer before deciding whether or not to implement the program is not possible.

While this is a real concern, this should not prevent evaluation of the effect of the program on the first cohort to be exposed to the program. While policy decisions will have to be taken in the meantime, knowing the answer at some point is surely better than never knowing it, which would be the case without evaluation. Moreover, obtaining short-term results, which may be used to get an indication of whether or not the program is likely to be effective, is often possible and may guide policy in the short run. For example, in the case of the evaluation of the HIV/AIDS teacher training program, an assessment was performed a few weeks after the program was started and while it was still ongoing. Students in the schools where the teachers were first trained were interviewed about whether the curriculum in their school covered HIV/AIDS and were administered a knowledge, attitude, and practice test. The preliminary results suggested that the program was indeed effective in increasing the chance that HIV/AIDS would be mentioned in class and in improving students' knowledge about HIV/AIDS and HIV prevention. These results could be communicated immediately to the policymakers.

The first results of an evaluation can also be combined with other results or with theory to provide an estimate of what the final impact of the program is expected to be. Obviously, one has to be cautious about such exercises and carefully outline what comes out of the evaluation results and what is the result of assumptions. One should set up programs so that long-run outcomes can be tracked that can then vindicate or invalidate predictions. For example, Miguel and Kremer (2004) combined their estimate of the impact of the deworming program on school participation with estimates of returns to education in Kenya to provide an estimate of the long-term impact on adult productivity, which they used to construct their cost-benefit estimates. They are also continuing to track the children exposed to deworming drugs to directly estimate the drugs' long-run effect.

Finally, delaying some expenditures may actually be worthwhile, given that we know so little about what works and what does not, especially if this can give us an opportunity to learn more. It is disconcerting that we do not know more about what works and what does not work in education, for example, after spending so many years funding education projects. On this scale, the two or three years needed for an evaluation, or even the many more needed to obtain information about the long-run

outcomes, seem a short period of time. It may delay some expenditures, but it will accelerate the process of learning how to make these expenditures usefully. The U.S. Food and Drug Administration (FDA) requires randomized evaluation of the effects of a drug before it can be distributed. Occasionally, the delay the FDA imposes on the approval of new drugs has created resentment, most recently among associations representing AIDS victims; however, randomized trials have played a key role in shaping modern medicine and have accelerated the development of effective drugs.

The Role that International Agencies Can Play

This section discusses current evaluation practices and the role that international agencies can play in improving these practices.

Current Practice

The foregoing examples show that obtaining convincing evidence about the impact of a program is possible by organizing pilot projects, taking advantage of the expansion of existing projects, or taking advantage of project design. While not all programs can be evaluated using these methods, only a tiny fraction of those that could potentially be evaluated actually are. Most international organizations require that a fraction of the budget be spent on evaluation. Some countries also make evaluation compulsory; for example, the Constitution of Mexico requires evaluation of all social programs. However, in practice, this share of the budget is not always spent efficiently; for example, evaluations may be subcontracted to untrained consultancy outfits that are given little guidance about what they should achieve. Worse, they are sometimes entrusted to organizations that have an interest in the outcome, in which case the evaluators have a stake in the results they are trying to establish.

When an evaluation is actually conducted, it is generally limited to a process evaluation, that is, the accounts are audited; the flows of resources are followed; and the actual delivery of the inputs is confirmed, for example, whether textbooks reached the school. In addition, qualitative surveys are used to determine whether beneficiaries actually used the inputs (did the teachers use the textbooks?) and whether there is prima facie evidence that the program beneficiaries were satisfied by the program (were the children happy?). Process evaluation is clearly essential and should be part of any program evaluation: if no textbooks were actually distributed, finding that the program had no impact would hardly be surprising. However, just observing the beneficiaries' reactions to a program can lead to misleading conclusions about its effectiveness. Some programs may, from all observations, seem like resounding successes, even if they did not achieve their objectives. The emphasis on process evaluation implies that, more often than not, when impact evaluations take place they are an afterthought and are not planned for at the time the program starts.

India's District Primary Education Program (DPEP), the largest World Bank–sponsored education program, is an example of a large program that offered the

potential for interesting evaluations, but whose potential on this count was jeopardized by the lack of planning. The DPEP was supposed to be a showcase example of the ability to go to scale with education reform (Pandey 2000). Case (2001) provides an illuminating discussion of the program and the features that make its evaluation impossible. The DPEP is a comprehensive program that seeks to improve the performance of public education. It involves teacher training, inputs, and classrooms. Districts are generally given a high level of discretion in how to spend the additional resources. Despite the apparent commitment to a careful evaluation of the program, several features make a convincing impact evaluation of the DPEP impossible. First, the districts were selected according to two criteria: low level of achievement, as measured by low female literacy rates, but high potential for improvement. In particular, the first districts chosen to receive the program were selected "on the basis of their ability to show success in a reasonable time frame" (Pandey 2000, p. 14). The combination of these two elements in the selection process indicates that any comparison between the level of achievement of DPEP districts and non-DPEP districts would probably be biased downward, while any comparison of improved achievement between DPEP and non-DPEP districts (difference-in-differences) would probably be biased upward. This has not prevented the DPEP from putting enormous emphasis on monitoring and evaluation: the project collected large amounts of data and commissioned numerous reports. However, the data collection process was conducted only in DPEP districts. These data can only be used to do before and after comparisons, which clearly do not make any sort of sense in an economy undergoing rapid growth and transformation. If researchers ever found a credible identification strategy, they would have to use census or national sample survey data.

The Political Economy of Program Evaluation

I have argued that the problems of omitted variable bias that randomized evaluations are designed to address are real and that randomized evaluations are feasible. They are no more costly than other types of surveys and are far cheaper than pursuing ineffective policies. So why are they so rare? Cook (2001) attributes their rarity in education to the postmodern culture in American schools of education, which is hostile to the traditional conception of causation that underlies statistical implementation. Pritchett (2002) argues that program advocates systematically mislead swing voters into believing exaggerated estimates of program impacts. Advocates block randomized evaluations, because they would reveal programs' true impacts to voters. Kremer (2003) proposes a complementary explanation, whereby policymakers are not systematically fooled but have difficulty gauging the quality of evidence, knowing that advocates can suppress unfavorable evaluation results. Program advocates select the highest estimates to present to policymakers, while any opponents select the most negative estimates. Knowing this, policymakers rationally discount these estimates. For example, if advocates present a study showing a 100 percent rate of return, policymakers might assume that the true return is 10 percent. In this environment, if randomized evaluations are more precise (because the estimates are on average

unbiased), there is little incentive to conduct randomized evaluations because they are unlikely to be high enough or low enough that advocates will present them to policymakers.

Under such circumstances, international organizations can play a key role by encouraging randomized evaluations and funding them. Moreover, if policymakers and donors can more readily identify a credible evaluation when examples are already available, which seems plausible, this role can actually start a virtuous circle by encouraging other donors to recognize and trust credible evaluation, and thus advocate the generation of such evaluations as opposed to others. In this way, international organizations can contribute to a climate favorable to credible evaluation and overcome the reluctance noted earlier. The process of quality evaluation itself would then be scaled up above and beyond what the organizations themselves could promote and finance.

What International Agencies Can Do

The foregoing discussion suggests a number of actions that international organizations could undertake to strengthen the role of evaluations.

Defining Priorities for Evaluation

Demanding that all projects be subject to impact evaluation is almost certainly counterproductive. Clearly all projects need to be monitored to ensure that they actually happened, and thus to make sure that the international organization is functioning properly, which is the main responsibility of the organization's evaluation department. Some programs simply cannot be evaluated using the methods discussed in this paper, for example, monetary policy cannot be randomly allocated. Even among projects that could potentially be evaluated, not all need an impact evaluation. Indeed, the value of a poorly identified impact evaluation is low, and its cost, in terms of credibility, is high, especially if, as argued later, international organizations should 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 to avoid a potentially large waste of money. The committee's responsibility would be to assess the ability to deliver reliable, causal estimates of the project. 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 for which it will promote impact evaluations. Organizations could also set up randomized evaluations in other areas when the opportunity occurs.

Setting up Autonomous Impact Evaluation Units

Given the scarcity of randomized evaluations, there may be some scope for setting up a specialized unit to encourage, conduct, and finance randomized impact evaluations and to disseminate the results. Such a unit would also encourage data collection and

the study of true natural experiments with program-induced randomization. As noted earlier, randomized evaluations are not the only way to conduct good impact evaluations: when randomization is infeasible, other techniques are available. However, such evaluations are conducted much more routinely, while randomized evaluations are much too rare given their value and the opportunities for conducting them. They also have common features and would benefit from a specialized unit with specific expertise. Because impact evaluation generates international public goods, the unit could finance and conduct rigorous evaluations in the key areas the organization identifies.

Setting up an autonomous unit would have several advantages. First, it would ensure that conducting evaluation is a core responsibility of a team of people. Second, this unit would be free of the fire-walling requirements that are necessary to make the evaluation divisions of international organizations independent, but make prospective evaluations difficult. For example, the director of the World Bank's Operations Evaluation Department reports directly to the board, and the department's teams are prevented from establishing close connections with the implementation team. This makes a prospective randomized evaluation essentially impossible. Third, randomized evaluation and nonrandomized evaluation should be clearly separated to avoid the "scaling down" effect caused by the political economy of evaluation.

Banerjee and He (2003) argue that the World Bank's decisions and reports have little impact on market decisions or on subsequent debates, that is, that the World Bank does not seem to have the role of a leader and promoter of new ideas that it could have. This may be in part because everybody recognizes that the World Bank, perhaps legitimately, operates under a set of complicated constraints, and that what justifies its decisions it not always clear. Credibility would require the Bank to be able to separate the results generated from randomized evaluation from the data reported by the rest of the organization. The results of studies produced or endorsed by the unit could be published separately from other World Bank documents.

Working with Partners

An evaluation unit would have a tremendous impact in terms of working with partners, in particular, NGOs and academics. For projects submitted from outside the unit, a committee within the unit, perhaps with the assistance of external reviewers, could receive proposals from within the international organizations or from outsiders and choose projects to support. The unit could also encourage the replication of important evaluations by sending out calls for specific proposals. Many NGOs would certainly be willing to take advantage of the opportunity to obtain funding. NGOs are flexible and entrepreneurial and can easily justify working with only some people, because they do not serve the entire population. The project could then be conducted in partnership with people from the unit or other researchers, especially academics, to ensure that the team has the required competencies. The unit could provide both financial and technical support for this project with dedicated staff and

researchers. Over time, based on the experience acquired, the unit could also serve as a more general resource center by developing and disseminating training modules, tools, and guidelines for randomized evaluation. It could also sponsor training sessions for practitioners.

Certifying and Disseminating Evaluation Results

Another role the unit could serve, after establishing a reputation for quality, is acting as a certifying body, clearinghouse, and dissemination agency. To be useful, evaluation results need to be accessible to practitioners within and outside the development agencies. A role of the unit could be to conduct systematic searches for all impact evaluations, assess their reliability, and publish the results in the form of policy briefs and in a readily accessible and searchable database. The database should include all the information needed to interpret the results (estimates, sample size, region and time, type of project, cost, cost-benefit analysis, caveats, and so on), as well as some rating of the validity of the evaluation and references to other related studies. The database could include both randomized and nonrandomized impact evaluations and clearly label the different types of evaluation. 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. This would help alleviate the publication bias problem, whereby evaluations that show no results are not disseminated. While academic journals may be uninterested in publishing the results of programs that failed, from the point of view of policymakers, this knowledge is as useful as knowing about projects that succeeded. Ideally, over time, the database would become a basic reference for organizations and governments, in particular, as they seek funding for their projects. This database could then jump-start a virtuous circle, with donors demanding credible evaluations before funding or continuing projects, more evaluations being done, and the general quality of evaluation work rising.

Conclusion: Using Evaluation to Build Long-Term Consensus for Development

Rigorous and systemic evaluations have the potential to leverage the impact of international organizations well beyond 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. Programs that have been shown to be successful can be adapted for use in other countries and can be scaled up within countries, while unsuccessful programs can be abandoned. By promoting, encouraging, and financing rigorous evaluations of the programs they support, as well as of programs others support, the international organizations could provide guidance to the international organizations themselves, as well as to other donors, governments, and NGOs in the ongoing search for successful programs, and thereby improve the effectiveness of

development aid. Moreover, by credibly establishing which programs work and which do not, the international agencies could counteract skepticism about the effectiveness of spending on aid and build long-term support for development. This is the opportunity to achieve real scaling up.

Notes

1. See http://www.unicef.org/programme/lifeskills/priorities/index.html.
2. The World Bank is not immune to recommending programs whose effectiveness has not been established. A publication by Narayanan (2000) lists a series of programs recommended by the World Bank, of which few have been evaluated (Banerjee and He 2003).
3. PROGRESA is so called from the Spanish acronym for Program for Health, Education, and Nutrition.
4. See http://www.ifpri.org/themes/progresa.htm.
5. To make this point precisely, one would need a full cost-benefit analysis of both programs to see whether the same improvement in human capital were achieved with the same expenditure. The paper on India does not yet include a cost-benefit analysis.
6. Galasso, Ravallion, and Salvia (2002) use the same technique to control for endogenous take-up of a subsidy voucher and training program in Argentina, and Banerjee and others (2003) use it to control for the fact that only two-thirds of the schools allocated to the treatment group actually received the remedial education teachers.
7. A representative from a large organization once objected to the idea that randomized evaluations could be taught and "were not nuclear physics." His answer was that "studying human beings is much more complicated than nuclear physics." This exactly makes the point that, unlike for physics, there are no general laws of human behavior, and therefore nothing general can be said.

References

The word "processed" describes informally reproduced works that may not be commonly available in libraries.

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

_____. 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): 533–75.

Angrist, Joshua, Eric Bettinger, Erik Bloom, Elizabeth King, and Michael Kremer. 2002. "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment." *American Economic Review* 92(5): 1535–58.

Ashenfelter, Orley, Colm Harmon, and Hessel Oosterbeek. 1999. "A Review of Estimates of the Schooling/Earnings Relationship, with Tests for Publication Bias." *Labour Economics* 6(4): 453–70.

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

Banerjee, Abhijit. 2002. "The Uses of Economic Theory: Against a Purely Positive Interpretation of Theoretical Results." Massachusetts Institute of Technology, Cambridge, Mass. Processed.

Banerjee, Abhijit, and Ruimin He. 2003. "The World Bank of the Future." *American Economic Review Papers and Proceedings* 93(2): 39–44.

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." Harvard University and Massachusetts Institute of Technology, Cambridge, Mass. Processed.

Banerjee, Abhijit, Shawn Cole, Esther Duflo, and Leigh Linden. 2003. "Remedying Education: Evidence from Two Randomized Experiments." Massachusetts Institute of Technology, Cambridge, Mass. Processed.

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

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Difference in Differences Estimates?" *Quarterly Journal of Economics* 119(1): 249–75.

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." University of California, Berkeley. Processed.

Buddelmeyer, Hielke, and Emmanuel Skoufias. 2003. *An Evaluation of the Performance of Regression Discontinuity Design on PROGRESA*. Discussion Paper no. 827. Bonn: Institute for Study of Labor.

Campbell, Donald T. 1969. "Reforms as Experiments." *American Psychologist* 244: 7–29.

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." Princeton University, Princeton, N.J. Processed.

Chattopadhyay, Raghabendra, and Esther Duflo. Forthcoming. "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India." *Econometrica*.

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

Cronbach, L. 1982. *Designing Evaluations of Educational and Social Programs*. San Francisco: Jossey-Bass.

Cronbach, L., S. Ambron, S. Dornbusch, R. Hess, R. Hornik, C. Phillips, D. Walker, and S. Weiner. 1980. *Toward Reform of Program Evaluation*. San Francisco: Jossey-Bass.

Cullen, Julie Berry, Brian Jacob, and Steven Levitt. 2002. "Does School Choice Attract Students to Urban Public Schools? Evidence from over 1,000 Randomized Lotteries." University of Michigan, Ann Arbor. Processed.

DeLong, J. Bradford, and Kevin Lang. 1992. "Are All Economic Hypotheses False?" *Journal of Political Economy* 100(6): 1257–72.

Duflo, Esther. 2001. "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment." *American Economic Review* 91(4): 795–813.

Duflo, Esther, and Michael Kremer. Forthcoming. "Use of Randomization in the Evaluation of Development Effectiveness." In *Proceedings of the Conference on Evaluating Development Effectiveness*. Washington, D.C.: World Bank, Operations Evaluation Department.

Duflo, Esther, Michael Kremer, and Jonathan Robinson. 2003. "Understanding Technology Adoption—Fertilizer in Western Kenya: Preliminary Results from Field Experiments." Massachusetts Institute of Technology, Cambridge, Mass. Processed.

Duflo, Esther, Pascaline Dupas, Michael Kremer, and Samuel Sinei. 2003. "Evaluating HIV/AIDS Prevention Education in Primary Schools: Preliminary Results from a Randomized Controlled Trial in Western Kenya." Harvard University and Massachusetts Institute of Technology, Cambridge, Mass. Processed.

Galasso, Emanuela, Martin Ravallion, and Agustin Salvia. 2002. "Assisting the Transition from Workfare to Work: A Randomized Experiment." World Bank, Development Research Group, Washington, D.C. Processed.

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

Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2003. "Teacher Incentives." Harvard University, Cambridge, Mass. Processed.

Glewwe, Paul, Michael Kremer, Sylvie Moulin, and Eric Zitzewitz. Forthcoming. "Retrospective Versus Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya." *Journal of Development Economics*.

Heckman, James, Lance Lochner, and Christopher Taber. 1998. "General Equilibrium Treatment Effects: A Study of Tuition Policy." Working Paper no. 6426. National Bureau of Economic Research, Cambridge, Mass.

Hsieh, Chang-Tai, and Miguel Urquiola. 2002. "When Schools Compete, How Do They Compete? An Assessment of Chile's Nationwide School Voucher Program." Princeton University, Princeton, N.J. Processed.

Imbens, Guido, and Joshua Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62(2): 467–75.

Kremer, Michael. 2003. Randomized Evaluations of Educational Programs in Developing Countries: Some Lessons." *American Economic Review Papers and Proceedings* 93(2): 102–15.

Krueger, Alan. 1999. "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics* 114(2): 497–532.

LaLonde, Robert. 1986. "Evaluating the Econometric Evaluations of Training with Experimental Data." *American Economic Review* 76(4): 604–20.

Meyer, Bruce D. 1995. "Natural and Quasi-Experiments in Economics." *Journal of Business and Economic Statistics* 13(2): 151–61.

Miguel, Edward, and Michael Kremer. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica* 72(1): 159–218.

Morduch, Jonathan. 1998. "Does Microfinance Really Help the Poor? New Evidence from Flagship Programs in Bangladesh." Princeton University, Princeton, N.J. Processed.

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

Pandey, Raghaw Sharan. 2000. *Going to Scale with Education Reform: India's District Primary Education Program, 1995–99*. Education Reform and Management Publication Series, vol. I, no. 4. Washington, D.C.: World Bank.

Pitt, Mark, and Shahidur Khandker. 1998. "The Impact of Group-Based Credit Programs on Poor Households in Bangladesh: Does the Gender of Participants Matter?" *Journal of Political Economy* 106(5): 958–96.

Pritchett, Lant. 2002. "It Pays to Be Ignorant: A Simple Political Economy of Rigorous Program Evaluation." *Journal of Policy Reform* 5(4): 251–69.

Rosenbaum, Paul R. 1995. "Observational Studies." In *Series in Statistics*. New York: Springer.

Shultz, T. Paul. Forthcoming. "School Subsidies for the Poor: Evaluating the Mexican PROGRESA Poverty Program." *Journal of Development Economics*.