Scaling Up and Evaluation[1]

Esther Duflo

Paper prepared for the ABCDE in Bangalore
May 21-22, 2003

Abstract

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 small scale would work in other contexts. Therefore, the very objective of scaling up implies that it is possible to learn from past experience.

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: The international agencies should thus have a key role in promoting and financing them. In doing this, they would achieve three important objectives: Improve the rates of returns on the programs they support; improve the rates of returns on the programs other policymakers support, by providing evidence on the basis of which programs can be selected; 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 there is considerable scope to expand the use of randomized evaluations. For a broad class of development programs (although not all of them), randomized evaluation can be used to overcome the problems often encountered when using evaluation practices. First, it discusses the methodology of randomized evaluation through several concrete examples, mostly drawn from India. It then discusses the potential of randomized evaluation as a basis for scaling up. Finally, it discusses current practices and the role international agencies can play in promoting and financing rigorous evaluations.

[1] I thank Francois Bourguignon, Angus Deaton and T. P. Schultz for extremely detailed and useful comments to a previous draft. The paper also benefited enormously from the collaboration with Michael Kremer on a related paper, prepared for the OED Biannual Evaluation Conference (Duflo and Kremer 2003). I thank Abhijit Banerjee, Edward Miguel and Martin Ravallion, for helpful comments. Finally, I am grateful to Nick Stern for asking two critical questions (How to generalize evaluation results? How can we learn quickly?), which greatly influenced the revision of this paper, even though this paper does not satisfactorily answer them. I am fully responsible for the content of this paper, which does not necessarily represent the view of the World Bank or any other agency. I gratefully acknowledge financial support from the Alfred P. Sloan Foundation.

"Scaling up" and "evaluation" are often presented as conflicting objectives, and, for most international development agencies, "going to scale" has to be given priority. 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".[2] The tradeoff is explicit in this heading: 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" -- the set of "effective" projects would be too small. The UNICEF "Skilled Based Health Education" website reports on ten case studies of "promising" school-based HIV/AIDS education programs, only one of which presents differences in outcomes between a treatment and a comparison group. These approaches are the programs that UNICEF can recommend be implemented on a national scale.[3]

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 it is possible to learn from past experience. 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,

[2] See, http://www.unicef.org/programme/lifeskills/priorities/index.html.
[3] The World Bank is not immune to recommending programs whose effectiveness has not been established: The publication Empowerment and Poverty Reduction: A Sourcebook (Narayanan 2000) lists a series of programs recommended by the World Bank, of which very few have been evaluated (Banerjee and He 2003).

governments, donors, and NGOs in their ongoing search for effective programs. Therefore, 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. It is true that a prospective impact evaluation may require postponing the national expansion of a program for some time. However, evaluation can be part of the backbone of a much larger expansion: That of the project on a much larger scale (if proved 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, which overcome the problems often encountered when using other evaluation practices to estimate program impact. 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, local government reforms, education, and health) 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 article proceeds as follows: In Section 1, I present the impact evaluation problem, the opportunities for evaluation, and discuss examples of evaluations, which will be drawn mostly from India. In Section 2, I discuss the potential of randomized evaluation as a basis for scaling up. In Section 3, I discuss current practices and the role international agencies can play in promoting and financing rigorous evaluations. Section 4 concludes.

The Methodology of Randomized Evaluation

The Evaluation Problem

Any impact evaluation attempts to answer an essentially counterfactual question: How would individuals who did not 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, give us a reliable estimate of the impact the program had on him, since many other things may have changed at the same time that the program was introduced. We can therefore not seek to obtain an estimate of the impact of the program on each individual. All we can hope for is to be able to obtain the average impact of the program on a group of individuals, by comparing them to a similar group who were not exposed to the program. The critical objective of impact evaluation is therefore to establish a credible *comparison group*, a group of individuals who *in the absence of the program* would have had outcomes similar to those who were exposed to the program. This group gives us an idea of what would have happened to the program group if they 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: Programs are placed in specific areas (for example, poorer or richer areas), individuals are screened for participation in the program (for example, on the basis of poverty, or on the basis of their motivation), and finally the decision to participate is often voluntary. For all of these reasons, those who were not exposed to a program are often not a good comparison group for those who were: Any difference between them could be attributed to two factors: pre-existing differences (the so called "selection bias"), and the impact of the program. Since we have no reliable way to estimate the size of the selection bias, we cannot decompose the overall difference into a treatment effect and a bias term.

To solve this problem, program evaluations typically need to be carefully planned in advance, in order to determine which group is a likely control group. One situation where the selection bias disappears is when the treatment and 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 programs 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. This random selection of treatment and comparison groups can happen in several circumstances: during a pilot project, because the program resources are limited; or because the program itself calls for random beneficiaries. In the next two subsections, we discuss examples of these different scenarios. There are also circumstances where a program was not randomly allocated, but where, due to favorable circumstances, a credible control groups nevertheless exists.

Prospective Randomized Evaluations

Pilot projects. Before a program is launched on a large scale, a pilot project, necessarily limited in scope, is often implemented. Randomly choosing the beneficiaries of the pilot can be done in most circumstances, since many potential sites (or individuals) are as deserving to be the places where the pilot takes place. The pilot can then be used, not only if the program is feasible (which is what most pilots are used for at the moment), but also when the program has the expected impacts. Job training and income maintenance programs were prominent examples of randomized evaluations. A growing number of such pilot projects are evaluated, often in collaboration between an NGO and academics (see, for example, Kremer 2003 for several references). To illustrate briefly how these studies can work in practice, I chose an example from India, analyzed in Banerjee and others (2001). This study evaluated a program where an Indian NGO (Seva Mandir) decided to hire a second teacher in the non-formal education centers they run in villages. Non-formal 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 increase performance by providing more individualized

attention to the children. By providing a female teacher, the NGO also hoped to make school more attractive for girls. Teacher attendance and child attendance were regularly monitored in program and comparison schools during 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 a school was closed: One-teacher schools are closed 39% of the time, whereas two-teacher schools are closed 24% of the time. Girl's attendance increased by 50%. However, there was no difference in test scores.

Carefully evaluated pilot projects form a sound basis for the decision to scale the project up. In the example just discussed, the two-teacher program was *not* implemented on a full scale by the NGO, on the ground that the benefits were not sufficient to outweigh the cost. The savings were used to expand other programs. Positive results, on the other hand, can help build a consensus for the project, which has the potential to be extended far beyond the scale that was initially envisioned. The PROGRESA program in Mexico is the most striking example of this phenomenon. 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, officials in the Mexican government made a conscious decision to take advantage of the fact that budgetary constraints made it impossible to reach the 50,000 potential 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 was made accessible to many different people, and a number of papers have been written on its impact (most of them are accessible on the IFPRI website). The

evaluations showed that it was effective in improving health and education: Comparing PROGRESA beneficiaries and non-beneficiaries, Gertler and Boyce (2001) show that children had about a 23% reduction in the incidence of illness, a 1-4% increase in height, and an 18% reduction in anemia. Adults experienced a reduction of 19% in the number of days lost due to illness. Shultz (2001) finds an average of 3.4% increase in enrollment for all students in grades 1 through 8; the increase was largest among girls who had completed grade 6, at 14.8%. 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% of the families in Mexico, and had a budget of US $800 million, or 0.2% of GDP (Gertler and Boyce 2001). It was subsequently expanded to urban communities and, with support from the World Bank, similar programs are being implemented in several neighboring Latin American countries. Mexican officials transformed a budgetary constraint into an opportunity, and made evaluation the cornerstone of subsequent scaling up. They were rewarded by the expansion of the program, and by the tremendous visibility that it acquired.

Replication, and e*valuation of existing projects*. A criticism often heard against the evaluation of pilot projects is that … they are pilot projects. This can create problems with the interpretation of the results: If the project is unsuccessful, it may be because it faced implementation problems in 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 situation, 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, any program is implemented in 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 unsuccessful) experiments in different contexts. This presents 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 are very different in 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 2003) evaluated a program of twice-yearly school-based mass treatment with inexpensive deworming drugs in Kenya, where the prevalence of intestinal worms among children is very high. Seventy-five schools were phased into the program in random order. Health and school participation improved not only at program schools, but also at nearby schools, due to reduced disease transmission. Absenteeism in treatment schools was 25% (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 returns to schooling, the estimates suggest extremely high rates of returns of the deworming intervention: The authors estimate that deworming increases the net present value of wages by over $30 per treated child at a cost of only $0.49. One of the authors then decided to examine whether these results generalized among pre-schoolers in urban India (Bobonis, Miguel and Sharma 2002). The baseline revealed that, although worm infection is present, the levels of infection were substantially lower than in Kenya (in India, "only" 27% of children suffer from some form of worm infection). However, 70% of children had moderate to severe anemia. The program was thus modified to include iron supplementation. The program was administered through a network of pre-schools in urban India. After one year of treatment, they found a nearly 50% reduction in moderate to severe anemia, large weight gains, and a 7% reduction in absenteeism among 4 to 6 year olds (but not for younger children). The results of the previous evaluation were thus by and large vindicated.[4]

[4] 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 was achieved with the same expenditure. At this point, the paper on India does not have a cost-benefit analysis yet.

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 implanted. It also may make it easier to evaluate the program in several sites at the same time, and thus alleviate some of the concerns about internal 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, Cole, Duflo and Linden (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 over 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 grade 2, 3 or 4 without having mastered the basic grade 1 competencies. Children who are identified as lagging behind are pulled out of the regular classroom for two hours a day to receive this instruction. Pratham wanted to evaluate the impact of this program, one of 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) of the 123 Vadodara government schools. In 2000-2001, the program was expanded to all the schools, but the 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 serve as the comparison group for Grade 3 students in schools that got the program in grade 4. 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 school switching groups. The program is thus conducted in several grades, in two cities, and with no school feeling that they are deprived of resources relative to others, since 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 an even stronger impact on the test scores of the 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 very similar across cities and child gender. 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, since the program was continued for a period of time, it was evaluated in two very 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, since the evaluation exercise does not require any modification of the design of the program.

Allocation to particular schools is often done by lottery, when some schools are oversubscribed. In some school systems in the U.S., students have the option of applying to "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. Since 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 what would have been expected and what a "naïve" comparison would have found: The results of children who attended a school of their choice are indeed better than that of those who do not, but this reflects the fact that the children who decided to change schools were highly motivated.

Voucher programs constitute another example of programs which often feature a lottery: The government 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)

evaluated a Colombian program in which vouchers for private schools were allocated by lottery, because of the limitation in the program's budget. Vouchers were renewable conditional on satisfactory academic performance. They compare lottery winners and losers. Lottery winners were 15-20% more likely to attend private school, 10% more likely to complete 8th grade, 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/college entrance exams. The benefits of this program to participants clearly exceeded the cost, which was similar to the cost 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 in other countries. However, because the randomization is part of the program design, rather than a deliberate attempt to make it possible to evaluate it, the data that makes evaluation possible is 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 (the Panchayat system) in India 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, for the administration of 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 (of council members and council chairpersons) 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 position be randomly allocated. Chattopadhyay and Duflo (2001) evaluated the impact in West Bengal of the reservation of the seats for women.

They collected data in 465 villages in 165 councils in one district, and they found that women tend to allocate more resources to drinking water and roads and less for education. This corresponds to the priorities expressed by men and women through their complaints to the Panchayat authorities. Before completing a second draft of this paper (Chattopadhyay and Duflo 2003), they collected the same data in a poor district of Rajasthan, Udaipur. They found that there, women invest more in drinking water, and less on roads, and that this corresponds again to the ordering of 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; Rajasthan is also one of the Indian States with particularly low female literacy), suggesting that the gender of the policymakers matters both in 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. They also evaluated the impact of reservation to scheduled castes, and found that a larger share of goods gets attributed to scheduled castes 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 do exist: Village-level census data is available for 1991, and will become available for 2001. The National Sample Survey Organization (NSSO) conducts large-scale detailed consumption and labor surveys every five years, with detailed data on outcomes. However, administrative barriers make this data very difficult to use for the purpose of evaluating this program: The census does not contain any information about which Panchayat a village belong to. The information about Panchayat reservation and composition is not centralized, even at the State level (it is available only at the district level). Likewise, the NSS contains no information about the Panchayat. This is an example where, at a relatively small cost, it would be possible to make available information useful to evaluate a very large program. It requires coordination of various people and various agencies, a task that the 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 which 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. Tremendous progress has been made, notably by labor economists. This article is not really the place to discuss them, and there are excellent technical and non-technical surveys of these techniques, their value as well as their limitations (see, for example, Angrist and Krueger 1999, 2001; Card 1999; and Meyer 1995). I only briefly mention here some of the techniques that are most popular with researchers.

A first strategy is to try to find a control group that is as "comparable" as possible to 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, i.e., 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 form a comparison group by picking people who have the same probability of being treated as those who actually got treated ("propensity score matching," Rosenbaum 1995). The challenge with this method, as in regression controls, is that it hinges on having identified all the potentially relevant differences between treatment and controls. In cases where the treatment is assigned on the basis of a variable that is not observed by the researcher (demand for the service, for example), this technique 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 be put in place, it is possible to compare the *growth* in the variables of interest between program and non-program regions (this is often called the "difference-in-differences" technique). It is often hard to judge whether the argument is good, however, and the identification

assumptions are justified. 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 a 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. And finally, when drawing inferences, one needs to take into account the fact that regions are often affected by time-persistent shocks, which may look like a "program effect" (Bertrand, Duflo and Mullainathan 2003). 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 by the fact that the allocation rule for the school is known (more schools were built in places with low initial enrollment rates), and by the fact that 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 very 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: Often, when policy changes are used to identify the effect of a particular policy, the policy change is itself endogenous to the outcomes they tried to affect, which makes identification impossible (see Besley and Case 2000).

Finally, the program rules often generate discontinuities that can be used to identify the effect of the program by comparing those who made it to those who "almost made it". For example, if scholarships are allocated on the basis of a certain number of points, it is possible to compare those just above to those just below the threshold. Angrist and Lavy (1999) used this technique (called regression discontinuity design [see Campbell 1969]) to evaluate the impact of class size in Israel. In Israel, a second teacher is allocated every time the class size would be above 40. 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, etc. Angrist and Lavy compared test score performances in schools just above and just below the threshold, and found that those just above the threshold have

significantly higher test scores than those just below, which can confidently be attributed to the class size, since it is very difficult to imagine that schools on both sides of the threshold have any other systematic differences. Discontinuities in program rules, when enforced, are thus source of identification. However, they often are NOT implemented, especially in developing countries. For example, researchers tried to use as a source of identification the discontinuity in Grameen bank (the flagship microcredit organization, in Bangladesh), which lends only to people who own less than one acre of land (Pitt and Khandker 1998). However, it turns out that *in practice*, Grameen bank lends to many people who own more than one acre of land, and that there is no discontinuity in the probability for borrowing at the threshold (Morduch 1998). In developing countries, it is likely to often be the case that rules are not enforced strictly enough to generate discontinuities that can be used for identification purposes.

Alternatives to randomized evaluation exist, and they are very 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 which need to accompany them. This suggests that, while a mix of randomized and non-randomized evaluation is necessary, there should be a commitment to run some randomized evaluations in international organizations.

Scaling Up and Randomized Evaluations

The previous section has shown 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 program effects. In this section, I discuss 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*, in order to evaluate the impact of a program, researchers must resort to before and after comparisons (when a baseline was conducted), or 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 not known (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 we have argued above, controlling for observable differences between treatment and control groups (through a regression analysis or through propensity score matching) will be correct for the bias only if it is known with certainty that beneficiaries and non-beneficiaries are 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 the resources.

These concerns have serious practical implications. Studies comparing experimental and non-experimental estimates with the same data show that the results from randomized evaluation can be quite different from those drawn from non-randomized evaluation. In a celebrated analysis of job training programs, LaLonde (1986) found that many of the econometric procedures and comparison groups used in program evaluations did not yield accurate or precise estimates, and that such econometric estimates often differ significantly from experimental results. Glewwe and others (2003) compared retrospective and prospective analyses of the effect of flip charts on test scores. Retrospective estimates using straightforward OLS regressions suggest that flip charts raise test scores by up to 20% of a standard deviation, robust to the inclusion of control variables; difference-in-difference estimates suggest a smaller effect of about 5% of a standard deviation, an effect that is still significant though sometimes only at the 10%

level. In contrast, prospective estimates based on randomized evaluations provide no evidence that flip charts increase test scores. These results suggest that using retrospective data to compare test scores seriously overestimates the charts' effectiveness. A difference-in-difference approach reduced but did not eliminate the problem and, moreover, it is not clear that such a difference-in-difference approach has general applicability. These examples suggest that OLS estimates are biased upward, rather than downward. This is plausible, since in a poor country with a substantial local role in education, inputs are likely to be correlated with favorable unobserved community characteristics. If the direction of omitted variable bias were similar in other retrospective analyses of educational inputs in developing countries, the effects of inputs may be even more modest than retrospective studies suggest. Some of the results are more encouraging: For example, Buddlemeyer and Skoufias (2003) used 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 ten out of twelve cases, and the two 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. There is a natural tendency for positive results 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, clearly 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 only select large estimates. For example, Ashenfelter, Harmon and Oosterberbeek (1999) show that there is 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 result that instrumental estimates of the returns to education are higher than ordinary least squares estimates. In contrast, randomized evaluations commit in advance to a particular comparison group: Once the work is done to conduct a prospective randomized evaluation, one just needs to make sure 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 will discuss below, it is important to put institutions in place to ensure negative results are systematically disseminated (such a system is already in place for medical trials results).

There are also several sources of bias that 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 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: Although 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) re-analysis of the Tennessee STAR class size experiment used exactly this method to deal with the fact that some parents had managed to re-allocate their children from "regular" classes to small classes.[5] 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 (e.g., 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, since people who anticipated benefiting more from the program may be more likely to take advantage of it. It may, however, be a group that the policymakers especially care about, since 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 analyzed by Banerjee, Jacob and Kremer (2001) increased school attendance and reduced drop out. 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 downwards. 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

[5] Galasso, Ravallion and Salvia (2002) use the same technique in order 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.

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, the study by Miguel and Kremer (2003) of the Kenyan de-worming program showed that children in treatment schools (and in schools near to 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, since 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, schools or communities), rather than across individuals within a unit is also often the only practical way to proceed. For example, it may be impossible to offer a program to some villagers and not others. 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 are 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 due to the non-random treatment assignment cannot either be signed or estimated.

Generalizing the Results of Evaluation: "We Are Not in the Mcdonald's Business"

Randomized evaluation thus can provide reliable estimates of treatment effects for the program and the population under study. In order to draw on these estimates to assess the prospect for the program to be scaled up, 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 generalize.

First, the experiment itself may have affected the treatment or the comparison samples: For example, provision of inputs might temporarily increase morale among beneficiaries and this could improve performance. While this would bias randomized evaluations, it would also bias fixed-effect or difference-in-difference estimates. As mentioned above, either the treatment or the comparison group may also be temporarily affected by being part of an experiment (these phenomena are called, respectively, *Hawthorne* and *John Henry* effects), but these effects are less likely to be present when the evaluations are conducted on a large scale and over a long enough time span. 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. It is, however, important to try to assess whether these effects are present. In his re-analysis of the project STAR 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, since 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 Columbian voucher program analyzed in Angrist and others (2002), which we described above, 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 the private schools). If this

program were to be implemented on a large scale, it may affect the functioning of the school system, and thus 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). To address these problems, we need randomized evaluation performed at the level of the "economy". This may be possible for programs such as voucher program, where the general equilibrium effects will plausibly take place at the level of the community, 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 ("We are not in the MacDonald's business," to use Nick Stern's phrase). 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 (e.g., Cronbach and others 1980, and Cronbach 1982; and 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 within some international organizations,[6] but it is important to 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 "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.

[6] 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.

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. Importantly, theory will help unpack 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, 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 programs that are justified by some well-founded theoretical reasoning should be evaluated in priority, 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. Since prospective evaluations need to be planned ahead of time, it is also often possible to design pilot programs in such a way that they help in answering a specific question, or testing a specific theory. For example, Duflo (2003) reports on a series of randomized evaluations conducted in Kenya with Michael Kremer and Jonathan Robinson. They were motivated by the general question: Why are there so few farmers in this region of Kenya who use fertilizer (only about 10% of them do), despite the fact that it seems to be profitable and it is widely used in other developing countries, as well as 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 returns were often in excess of 100%). They then conducted a series of programs to answer the following 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 them, they implemented several programs: First, they randomly selected farmers to participate in the field trials, and followed their adoption subsequently, as well as that of a comparison group. Second, they also followed

adoption of 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 also monitored subsequent adoption. These questions are very 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, their answer also helped the NGO develop a school-based agricultural extension program which has a chance to be effective and cost effective. A pilot version of this program is currently being evaluated.

Theory and existing evidence can thus be used to design informative replication experiments, and to sharpen the prediction from these experiments. Rejection of these predictions should then be taken seriously, and will inform the development of the theory. In the first section, we gave several examples of replication. 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 a similar program. Some of these programs have included a randomized evaluation (for example, the PRAF program in Honduras), and are currently being evaluated.

It is also worth nothing that it is possible to use the exogenous variation created by the randomization to help identify a structural model. Attanasio, Meghir and Santiago (2001) and Berhman, Sengupta and Todd (2002) are two example of this exercise using the PROGRESA data to make some prediction of the possible effect 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: In fact, they generate data and variation which can help in identifying some aspects of these theories.

The Feasibility of Randomized Evaluation

As we 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 the 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. In fact, 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 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 math is comparable to the effect of the remedial education program, but the cost is much smaller. Even keeping constant the budget of process evaluation, 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 to the amount of money saved by avoiding the expansion of ineffective programs. This suggests that

randomized evaluation should be financed by international organizations, a point to which we will return below.

Political economy concerns sometimes make it difficult to not implement the program in the entire population, especially when its success has already be demonstrated (for example "opportunidades", the urban version of PROGRESA, will not start with a randomized evaluation, because of the strong opposition to delaying access to it to some people). 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 we have argued above, the evaluations are not financed by a loans, but by grants, this may make it easier to convince partners of its usefulness, especially if it makes it possible for the country to expand a program. An example of such explicit partnership is a study on the effectiveness of HIV/AIDS education, currently being conducted in Kenya (Duflo, Dupas, Kremer and Sinei 2003). With support from UNICEF, the government of Kenya has put together a teacher-training program for HIV/AIDS education. For lack of funds, the coverage of the program had remained very partial. The Partnership for Child Development, with grants from the World Bank, is funding a randomized evaluation of the teacher-training program. ICS, a Dutch NGO, is organizing training sessions, with facilitators from the Kenyan government. The evaluation has made it possible to expand training to 540 teachers in 160 schools, which would not have been possible otherwise. The randomization was not a ground for rejection of the program by the Kenyan authorities. On the contrary, at a conference organized to launch the program, Kenyan officials explicitly appreciated the opportunity the evaluation gave them to be at the forefront of efforts to advance knowledge on this question. The example of PROGRESA showed that government officials recognized the value of randomized evaluation, and were actually prepared to pay for it. The very favorable response to PROGRESA and the subsequent endorsement of the findings by the World Bank will certainly have an impact on 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 can therefore not hope to reach all the people they target. Randomized allocation is often perceived as a fair way to allocate sparse resources. In addition, their members are often very entrepreneurial, and willing to evolve in response to new information. NGOs tend to welcome information on the effectiveness of their programs, even if to find out that they are ineffective. For these reasons, many NGOs are willing to participate to randomized evaluations of their programs. For example, the collaboration between the Indian NGO Pratham and MIT 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 partnership 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 between researchers (from these organizations, or from academia) and *bona fide* NGOs.

The Timing of Evaluation and Implementation

Prospective evaluations do take time: Convincing studies often go on for two or three years. It takes even longer to obtain long-term impact of the program, which can be very important, and differs from the short run impact. 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 very long. In these cases, it is not possible to wait for the answer before deciding whether or not to implement the program.

While this is a real concern, this should not prevent the setting up of the evaluation on the first cohort to be exposed to the program: While it is true that policy decisions will have

to be taken in the meantime, it is surely better to know the answer at some point rather than never, which would be the case without evaluation. Moreover, it is often possible to obtain short-term results, which may be used to get an indication of whether or not the program has a chance to be effective, 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 HIV/AIDS was present in the curriculum in their school, and were administered a knowledge, attitude, and practice test. The preliminary results suggest that the program was indeed effective in increasing the chance that HIV/AIDS is 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 result 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 very careful about these exercises, and carefully outline what comes out of the evaluation results, and what is the result of assumptions. One should set up programs to be able to track long run outcomes, which can then vindicate or invalidate these predictions. For example, Miguel and Kremer (2003) combined their estimate of the impact of the de-worming program on school participation on 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 de-worming drug to directly estimate their 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 very 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 fact that an evaluation takes two or three years (or even many more to obtain information about the long run outcomes) seems a very short period of time. It may delay some expenditures, but it will accelerate the process of learning how to make these expenditures usefully. The FDA requires randomized

evaluation of the effects of a drug before it can be distributed. Occasionally, the delay it imposes on the approval of new drugs has created resentment (most recently, among associations representing AIDS victims). However, there is little doubt that randomized trials have played a key role in shaping modern medicine, and that they have accelerated the development of effective drugs.

The Role That International Agencies Can Play

Current Practice

The examples discussed above show that it is possible to obtain convincing evidence about the impact of a program by organizing pilot projects, taking advantage of expansion of existing projects, or taking advantage of project design. While not all programs can be evaluated using these methods, a very small fraction of those who could potentially be are. Most international organizations require that a fraction of the budget be spent on evaluation. Some countries also make evaluation compulsory (for example, evaluation of all social programs is required by the Constitution in Mexico). However, in practice, this share of the budget is not always spent efficiently: Evaluations get subcontracted to untrained consultancy outfits, with little guidance about what they should achieve. Worse, they are sometimes entrusted to organizations that have an interest in the outcome, so 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: Accounts are audited, the flows of resources are followed, the actual delivery of the inputs is confirmed (for example, did the textbooks reach the school?) and qualitative surveys are used to determine whether the inputs were actually used by their beneficiaries (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 also be part of any program evaluation: If no textbooks were actually distributed, finding no impact of the program is not going to be surprising. However, just observing the beneficiaries' reactions to a program can lead to very misleading conclusions about its effectiveness: Some programs

may, by 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, impact evaluations, when they take place, are an afterthought, and are not planned starting with the inception of the program.

The District Primary Education Program, the largest World Bank sponsored education program, implemented in India, is an example of a large program that offered the potential for very interesting evaluations, but whose potential on this count was jeopardized by the lack of planning. DPEP was supposed to be one showcase example of the ability to "go to scale" with education reform (Pandey 2000). Case (2001) gives an illuminating discussion of the program and the features that makes its evaluation impossible. DPEP is a comprehensive program seeking 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 DPEP impossible. First, the districts were selected according to two criteria: low *level* of achievement (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, quoted in Case 2001). The combination of these two elements in the selection process makes clear that any comparison between the level of achievement of DPEP districts and non-DPEP districts would probably be biased downwards, while any comparison between improvement of achievement between DPEP and non-DPEP districts ("differences-in-differences") would probably be biased upwards. This has not prevented the DPEP from putting enormous emphasis on monitoring and evaluation: Large amounts of data were collected, and numerous reports were commissioned. However, the data collection process was conducted *only in DPEP districts*! This data can only be used to do before/after comparisons, which clearly do not make any sort of sense in an economy undergoing rapid growth and transformation. If a researcher ever found a credible identification strategy, he or she would have to use census or National Sample Survey (NSS) data.

The Political Economy of Program Evaluation

We have argued that the problems of omitted variable bias which 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. Then, why are they so rare? Cook (2001) attributes their rarity in education to the post-modern culture in American education schools, which is hostile to the traditional conception of causation that underlies statistical implementation. Pritchett (2003) argues that program advocates systematically mislead swing voters into believing exaggerated estimates of program impacts. Advocates block randomized evaluations since they would reveal programs' true impacts to voters. Kremer (2003) proposed a complementary explanation, where 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 opponent selects the most negative estimates. Knowing this, policymakers rationally discount these estimates. For example, if advocates present a study showing a 100% rate of return, the policy maker might assume the true return is 10%. In this environment, if randomized evaluations are more precise (because the estimates are on average unbiased), there is little incentive to conduct randomized evaluations: They are unlikely to be high enough or low enough that advocates will present them to policymakers.

In this world, an international organization can play a key role by encouraging randomized evaluations and funding them. Moreover, if it becomes easier for policymakers and donors to identify a credible evaluation when there are already examples (which seems plausible), this role can actually start a virtuous circle, by encouraging other donors to recognize and trust credible evaluation, and thus advocate to generate such evaluation as opposed to others. In this way, they can contribute to a "climate" favorable to credible evaluation, and thus overcome the reluctance that we

mentioned above. The process of quality evaluation itself would then be scaled up above and beyond what the organizations can themselves promote and finance.

The Role International Agencies Can Play

The discussion in the preceding two sections suggests what international organizations could do to strengthen the role of evaluations.

Defining priorities for evaluation. It is almost certainly counter-productive to demand that *all projects* be subject to impact evaluation. Clearly, all projects need to be monitored to make sure that they actually happened to make sure the organization is properly functioning, which is the main responsibility of the Evaluation Department. Some programs can simply not be evaluated with the methods discussed in this paper: Monetary policy cannot be randomly allocated, for example. Even among projects which could potentially be evaluated, not all need an impact evaluation. In fact, the value of a poorly identified impact evaluation is very low, and its cost, in terms of credibility, is high, especially if international organizations, as we argue below they 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 of the project 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 where they will promote impact evaluations. They could also set up randomized evaluation 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 disseminate the results. Such a unit would also encourage data collection and study of true "natural experiments" with

program-induced randomization. As we mentioned above, randomized evaluations are not the only way to conduct good impact evaluations: When randomization is not feasible, other techniques are available. However, such evaluations are conducted much more routinely, while randomized evaluations are much too rare, in view of their value and the opportunities for conducting them. They also have common features, and would benefit from a specialized unit with specific expertise. Since impact evaluation generates international public goods, the unit could finance and conduct rigorous evaluations in the key areas identified by the organizations.

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 firewalling requirements that are necessary to make the evaluation divisions of the international organization independent, but make prospective evaluations difficult. For example, the director of the Operations Evaluation Department (OED) of the World Bank reports directly to the Board, and the OED teams are prevented from establishing close connections with the implementation team: This makes a prospective randomized evaluation essentially impossible. Third, there should be a clear separation between randomized evaluation and non-randomized evaluation, to avoid the "scaling down" effect due to 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: 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 it is not always clear what justifies their decisions. 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.

Work with partners. As previously discussed, such a 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 (potentially with the assistance of external reviewers) could receive proposals from within the Bank or outsiders, and choose projects to support. It could also encourage 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, entrepreneurial, and can easily justify working with only some people, since they do not have the vocation to serve the entire population. The project could then be conducted in partnership with people from the unit or other researchers (academics, in particular), to ensure that the team has the required competencies. It could provide both financial and technical support for this project, with dedicated staff and researchers. Over time, on the basis of the experience acquired, it could also serve as a more general resource center, by developing and diffusing training modules, tools, and guidelines (survey and testing instruments, software for data entry and to facilitate randomization – similar in spirit to tools produced by other units in the World Bank) for randomized evaluation. It could also sponsor training sessions for practitioners.

Certify and disseminate evaluation results. Another role the unit could serve, after establishing a reputation for quality, is that of a certifying body and "clearinghouse" and dissemination agency. In order to be useful, evaluation results need to be accessible to practitioners, within and outside 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 searchable database. The database should include all the information useful to interpret the results (estimates, sample size, region and time, type of project, cost, cost-benefit analysis, caveats, etc.), as well as some rating of the validity of the evaluation, and reference to other related studies. The database could include both randomized and non-randomized impact evaluations, satisfying some criteria, 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" (or "drawer") problem, whereby evaluations which showed no results are not

disseminated; academic journals may not be interested in publishing results of programs that failed, but from the point of view of policymakers, this knowledge is as useful as knowing about projects that succeeded. Comparable requirements are placed on all federally funded medical projects. 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 kick 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 simply their ability to finance programs. Credible impact evaluations are international public goods: The benefits of knowing that a program works or does not work extend well beyond the organization or the country implementing the program. Programs that have been shown to be successful can be adapted for use in other countries and scaled up within countries, while unsuccessful programs can be abandoned. Through promoting, encouraging, and financing rigorous evaluations (such as credible randomized evaluations) of the programs they support, as well as of programs supported by others, the international organizations can provide guidance to the international organizations themselves, as well as other donors, governments, and NGOs in the ongoing search for successful programs, and thus improve the effectiveness of development aid. Moreover, by credibly establishing which programs work and which do not, the international agencies can counteract skepticism about the possibility of spending aid effectively and build long-term support for development. This is the opportunity to achieve a real "scaling up".

References

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.

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: 1277-1366.

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

Angrist, Joshua, and Victor Lavy. 1999. "Using Maimonides' rule to estimate the effect of class size on scholastic achievement." *Quarterly Journal of Economics* 114 (2): 533-575.

Ashenfelter, Orley, Colm Harmon, and Hessel Oosterbeek. 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." Mimeo, Inter-American Development Bank.

Banerjee, Abhijit. 2002. "The Uses of Economic Theory: Against a Purely Positive Interpretation of Theoretical Results." Mimeo, MIT.

Banerjee, Abhijit, Shawn Cole, Esther Duflo, and Leigh Linden. 2003. "Remedying Education: Evidence from Two Randomized Experiments." Mimeo, MIT.

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." Mimeo, Harvard-MIT.

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

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2003. "How Much Should We Trust Difference in Differences Estimates?" forthcoming in *Quarterly Journal of Economics*.

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

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

Buddlemeyer, Hielke, and Emmanuel Skofias. 2003. "An Evaluation on the Performance of Regression Discontinuity Design on PROGRESA." Institute for Study of Labor, Discussion Paper No. 827.

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

Card, David. 1999. "The causal effect of education on earnings." In Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics, Vol. 3A*. Amsterdam: North Holland: 1801-63.

Case, Anne. 2001. "The primacy of education." Mimeo, Princeton University.

Chattopadhyay, Raghabendra, and Esther Duflo. 2001. "Women as Policy Makers: Evidence from a India-Wide Randomized Policy Experiment." NBER Working Paper # 8615.

Chattopadhyay, Raghabendra, and Esther Duflo. 2003. "Women as Policy Makers: Evidence from a India-Wide Randomized Policy Experiment." Mimeo, MIT.

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

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." Mimeo, University of Michigan.

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. 2003. "Poor but Rational?" Mimeo, MIT.

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", Mimeo, Harvard-MIT.

Duflo, Esther, and Michael Kremer. 2003. "Use of Randomization in the Evaluation of Development Effectiveness." Mimeo, MIT.

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

Gertler, Paul J., and Simone Boyce. 2001. "An experiment in incentive-based welfare: The impact of PROGRESA on health in Mexico." Mimeo, University of California, Berkeley.

Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2003. "Teacher Incentives." Mimeo, Harvard University.

Glewwe, Paul, Michael Kremer, Sylvie Moulin, and Eric Zitzewitz. 2003. "Retrospective vs. prospective analyses of school inputs: The case of flip charts in Kenya." forthcoming in *Journal of Development Economics.*

Heckman, James, Lance Lochner, and Christopher Taber. 1998. "General Equilibrium Treatment Effects: A Study of Tuition Policy." NBER Working Paper #6426.

Hsieh, Chang-Tai, and Miguel Urquiola. 2002. "When Schools Compete, How Do They Compete? An assessment of Chile's nationwide school voucher program." Mimeo, Princeton University.

Imbens, Guido, and Joshua Angrist. 1994. "Identification and estimation of local average treatment effects." *Econometrica* 62 (2): 467-475.

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

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): 604-620.

Meyer, Bruce D. 1995. "Natural and quasi-experiments in economics." *Journal of Business and Economic Statistics* 13 (2): 151-161.

Miguel, Edward, and Michael Kremer. 2003. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." forthcoming in *Econometrica*.

Morduch, Jonathan. 1998. "Does microfinance really help the poor? New evidence from flagship programs in Bangladesh." Mimeo, Princeton University.

Narayanan, Deepa, ed. 2000. *Empowerment and Poverty Reduction: A Sourcebook*. Washington DC: The 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, Volume I, No. 4*. Washington DC: 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): 958-996.

Pritchett, Lant. 2003. "It Pays to be Ignorant: A Simple Political Economy of Rigorous Program Evaluation." forthcoming, *Journal of Policy Reform*.

Rosenbaum, Paul R. 1995. "Observational studies." In *Series in Statistics*. New York; Heidelberg; London: Springer.

Shultz, T. Paul. 2001. "School Subsidies for the Poor: Evaluating the Mexican PROGRESA Poverty Program." forthcoming, *Journal of Development Economics*.