28101

Scaling up and Evaluation[1]

Esther Duflo

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

Abstract

This paper discusses the key 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 policy makers support, by providing evidence for the basis on 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 makes the case that the best way to estimate the impact of a broad class of development programs is through a randomized impact evaluation. First, it discusses the methodology of randomized evaluation through several concrete examples, mostly drawn from India. It then reviews the shortcoming of the current evaluation practices, before addressing the most frequent criticisms against randomized evaluation. Finally, it makes some suggestions about the role that international organizations could play in coordinating the accumulation of international knowledge on development projects: defining priorities for evaluation, fund and conduct randomized evaluation through an independent unit, work with partners, and create a data base of evaluation results.

[1] PRELIMINARY DRAFT, FOR CIRCULATION AT THE ABCDE CONFERENCI thank Ted Miguel, and, in particular, two reviewers for extremely detailed and useful useful comments. I also thank Abhijit Banerjee, Francois Bourguignon, Anne Case, Angus Deaton and Michael Kremer for numerous helpful discussions on the issues discussed in this paper.

For many international development agencies, "going to scale" is the absolute priority, even for programs whose effectiveness has not been fully established. UNICEF, for example, list 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" web site 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]

Understandably, for many practitioners and policy makers, time seems too short to wait for the "perfect" answer before rolling out programs. Dealing with situations of urgency, they can become impatient with those who want to take the time to implement rigorous evaluations of programs, and they need to act on the basis of whatever information is available. The situation is different for international organizations, except for the few cases where they deal with catastrophic crisis situations. These organizations have long run and global objectives (e.g., "A World Free of Poverty"). Their operations everywhere will benefit from the knowledge gained by evaluating one program. Knowledge about the effectiveness of a program is an international public good (it will benefit other countries), which these organizations are the best placed to provide. The time and money gained by avoiding replicating ineffective programs, and expanding those that have been shown to be successful could be enormous.

[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).

In this article, I argue that rigorous and systematic impact evaluations, and in particular randomized evaluations, far from being a barrier to scaling up, have the potential to leverage the impact of international organizations well beyond their ability to finance programs. Credible impact evaluations are international public goods: The benefits of knowing that a program works or does not work extends well beyond the organization or the country implementing the program.[4] Specifically, by encouraging and financing evaluation of the programs they support, as well as others, international organizations can achieve several objectives. First, they will provide guidance to themselves, other donors, country governments, and NGOs in their search for programs. Programs that have been shown to be effective can be replicated in other countries, and programs that have repeatedly shown to be ineffective can be abandoned. This should significantly improve the effectiveness of development budgets. Of course, this will not free policy makers from the necessity of making guesses of what may work and what may not, and will not prevent failures: The effectiveness of programs, and therefore of evaluations is in part context dependent. However, there could be no "scaling up" without the ability for programs to be replicated in different contexts. Theory needs to be used to guide the choice of what projects can work, and how the effectiveness of one project can be expanded to another, and evaluations are needed to test the theory. Second, by practicing and proponing these methods, they may raise the standards of acceptable evaluations upheld by other donors, thus multiplying the first effect. Last, improved rates of returns on development projects may increase the willingness to finance them, and increase overall aid budgets around the world.

A prospective impact evaluation may require postponing the national expansion of a program for some time (we will argue below that in most cases it will not postpone it for very long, given limited budget and implementation capability at the beginning of a program). In cases where the outcomes of interest will not be available before many years (such as the adult earnings of children exposed to an education program), setting up

[4] As we will see below, the benefits of a credible evaluation are often negative for the person who promotes the program.

a randomized treatment and comparison group can still be done, even if the final results are not available before the program is fully expanded. 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.

Of course, not all programs can be evaluated in this way: Monetary policy or democracy cannot be randomly allocated to one part of the country and not the other. This paper does not argue that *all* programs should be evaluated in this way, and offers no guidance for the evaluation of systems of macroeconomic programs (though it recognizes their importance!). Programs that are targeted to individuals or local communities (such as sanitation, local government reforms, education, and health) are likely candidates for credible impact evaluations, and this paper argues that, for these programs, international institutions should promote evaluations.

This article proceeds as follows: In Section 1, I present the impact evaluation problem, the opportunities for evaluation, and discuss examples of evaluations. I then contrast these with current evaluation practices. Examples will be drawn mostly from India. In Section 2, I will discuss possible reasons for the resistance to prospective impact evaluation that was encountered in many quarters. I will first discuss the arguments that are opposed to impact evaluation, and the extent to which they are valid concerns. Having argued that these arguments, while raising useful concerns, are not sufficient to explain how rare systematic evaluations are, I will then turn to other possible explanations. In Section 3, I will offer some suggestions on how international institutions can put evaluation at the core of their development strategy.

1. Evaluation: Best Practice

1.1 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 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.

1.2 Prospective Randomized Experiments

1.2.1 *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 pilot are used for at the moment), but also whether 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 et al. (2000). This study evaluated a program where an Indian NGO (Seva Mandir) decided to hire a second teacher in 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 during the entire duration of the project. The program reduced the number of days a school was closed: One-teacher schools are closed 44% of the time, whereas two-teacher schools are closed 39% of the time. Girl's attendance increased by 50%. However, there were no differences 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 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 of 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 web site). 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.

1.2.2 *Replication, and Evaluation 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 (this is called the Hawthorne effect, when the pilot treatment group changes their behavior as a result of being watched in an experiment, or a John Henry effect, when it is the comparison group's behavior that is affected). Even if randomization of the comparison and the treatment group ensure the internal validity of the estimates, the *external* validity of pilot projects may be limited (that is, the results cannot necessarily be generalized to other contexts); first, the impact of the program may have been modified by the evaluation itself; and second, the special conditions in which they were implemented may make it difficult to replicate them.

A first answer to some of these concerns is to replicate successful (and potentially unsuccessful as well) 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 later 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 4-6 year olds (but not for younger children). The results of the previous evaluation were thus by and large vindicated.[5]

A second answer is to evaluate programs that have already shown their potential to go to 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 evaluations 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 one 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.

[5] To make this point precisely, one would need a full cost-benefit analysis of both programs, to see whether the same improvement in human capital was achieved with the same expenditure. At this point, the paper on India does not have a cost-benefit analysis yet.

After two years, the program showed a very large impact on the test scores of all children (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 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.

1.3 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 selection into those schools.

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, et al. (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. 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 than 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.

1.4 Other Natural Experiments

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 well 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 at 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 as being treated than 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 important not to take this assumption for granted, however. This identification assumption cannot be tested, but 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, 2002). 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 parallel 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, 2001).

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 score 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 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.

2. The Resistance to Evaluation

2.1 Current practice

The various 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 projects design. Yet, such evaluations are few and far between. Most International organizations requires that a fraction of the budget be spent on evaluation (the World Bank requires 1%). 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. Worst, 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 where 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, as a part of program evaluation: If no textbooks were actually distributed, finding no impact of the program is not going to be surprising. However, by 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 a resounding success, 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. Researchers then resort to comparisons that are less than perfect: Before and after comparisons (when a baseline was conducted), or comparisons between beneficiaries and communities which, 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 not very informative. The data collection is often as expansive as for a randomized evaluation, but the inferences are biased. Controlling for observable differences between treatment and control groups (through a regression

analysis or through propensity score matching) will in general not be sufficient, unless it is known with certainty that beneficiaries and non-beneficiaries are comparable conditional on these characteristics, which 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. Glewwe et al. (2000) illustrate the bias of such analyses in the case of flipcharts. While a retrospective study comparing test scores between schools with and without flipcharts shows a positive impact, robust to the inclusion of control variables, a randomized experiment in the same region showed no impact of flipcharts on learning whatsoever.

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, was the District Primary Education Program, the largest World Bank sponsored education program, implemented in India. 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). It is rare to see advocacy being so candid . . . . 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 NSS data.

2.2 The Limitations of Randomized Evaluation

If randomized evaluations are the most convincing way to produce credible estimates of program impacts, and if those estimates are needed to form a credit steppingstone for scaling up, why do they remain so rare? Part of the reason is that they are strongly resisted in some quarters. Cook (2001) provides an excellent discussion of the arguments against randomized evaluation in education schools, and the present discussion draws from this article. Some of these arguments point to real limitations of randomized evaluations, which need to be understood, while some are less convincing.

2.2.1 *Evaluations are too expensive and take too much time: Policymakers cannot wait for them.*

The main cost of evaluation is the cost of data collection. While it does have a cost, it is no more expensive than the cost of collecting any other data. As we discussed previously, the M&E system of DPEP is largely a waste of money: It would certainly have been possible to set up a randomized evaluation of a part of DPEP for a fraction of the cost. 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". Even keeping constant the budget of process evaluation, a reallocation of the money that is currently spent on low quality impact 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. It is important to reiterate that the benefits of an evaluation extend far beyond the country or the organization that implemented the program. For those who implemented the program, the rate of return of the evaluation may well turn out to be negative, especially for small NGOs, who do not have the ability to expand. Below, we will also discuss political economy arguments which are likely to make the rates of returns to randomized evaluations negative. This suggests that randomized evaluation should be financed by international organizations, a point to which we will return below.

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 on test scores, but no long run impact, which they attribute to practices of "teaching to the test". This generates two potential problems. First, it may not be possible to delay the program for very long. Second, policymakers "cannot wait to get the perfect answer," they have to implement policies. The first argument is often not relevant, since those in the comparison group in the cohort that was initially not exposed to the program remain unexposed: For example, the long run impact of the teacher incentive program was measured after the program was stopped. It could also have been measured after the program was expanded to all subsequent cohorts. The second point, like the argument about cost, fails to recognize the difference in objectives between policymakers at the country level and international agencies: While policymakers have short horizons, international agencies, except in situations of urgency, should adopt a longer horizon and a broader perspective. It is surely preferable to get the right answer about the long run impact of the program at some point in the future than never, which is the case without evaluation. All too often, the previous decade (or decades) is referred to as the "wasted decade for…" (last time I heard this expression, the 80s and 90s were the wasted decades for education, the words of a UNICEF representative). It is of course very difficult to

know whether the development aid during a decade was actually wasted, since we do not know what would have happened in its absence. What is clear is that these decades were largely wasted in terms of our learning experience. 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 expenditure, but it will accelerate the process of learning how to make these expenditures usefully. If programs had been evaluated more systematically since the World Bank and UNICEF were constituted, one would hope that more would have been known by now. The analogy with drug development is useful. 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.

Furthermore, not all projects need to be evaluated (and many cannot be), and in the short run, not all projects will be justified by evidence arising from a randomized evaluation. International organization may want or need to finance those, for political reasons, or when they face an urgent situation. There is no reason why this should be impossible, as long as a transparent distinction is maintained between these projects and those which are evaluated, or supported, as a result of an evaluation.

2.2.2 *Randomization is often not feasible in practice*

Another argument against randomized evaluations is that they are often not feasible, because of opposition by the main actors. World Bank country officials are often against recommending randomized evaluations because “this is the country’s money”, and “ the government will never agree to it”. Political economy concerns make it difficult to not implement the program in the entire population (for example “*opportunidades*”, the urban

version of PROGRESA, will not start with a randomized evaluation, because of the strong opposition to delaying access to it to some people). Quite apart from the fact that international institutions are understandably reluctant to impose their will on governments, if people participating into the project are not opposed to the principle, even if randomization is agreed on, there will be deviations from it in practice, and those will jeopardize the results.

This objection can be tackled at several levels. First, according to Cook (2001) the same argument is very frequently heard in education schools, where teachers and principals are supposed to be against randomization. He attributes this to the lack of institutional support for randomization, and argues that when the school district has endorsed a study involving randomization, school teachers participated without difficulty. Opposition to randomization is likely to falter in an environment where it has strong support. Second, if the evaluations are not financed by a loans, but by grants, as we have already argued, (because of the externalities they generate), 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 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 et al., 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 school teachers 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 potential is enormous.

2.2.3 *Randomized evaluations are prone to biases and misleading inferences*

The reluctance of people to be involved in a randomized evaluation (especially in the comparison group) can jeopardize the quality of any evaluation, however well conceived in practice. This can introduce serious biases, which, for some, undo the advantages of randomized evaluations over other evaluation techniques. These sources of biases exist, and do need to be taken seriously. 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. 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), which is often the group the policy maker cares most about (since they will be those affected by the future policy).

A second possible source of bias is differential attrition in the treatment and the comparison groups: Those who benefit from the program may be less likely to move or otherwise drop out of the sample that 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 from dropping out by the program are the weakest in the class, the comparison between the test scores of children in treatment and control schools may be biased 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 et al., 2003), an attempt was made to track down *all* children and administer the test to them, even if they had

dropped out of school. Only children who had left for their home village were not tested. As a result, the attrition rate remained relatively high, but was not different in the treatment and comparison schools, and do 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. This reinforces the point that

these evaluations need to be conducted carefully by experienced people, and cannot be left to untrained consultants.

2.2.4 *Randomized evaluations have little or no external validity*

The most important concern against randomized evaluations is that, even if precautions are taken to circumvent any possible biases and thus insure *internal validity* (unbiased estimate of the effect of the program *in this group*), the results may not generalize to another context or another population. In its most extreme form (e.g., Cronbach et al, [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 learnt about schools *in general*. This discourse has made its way within some international organizations,[6] even though it seems contradictory with 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 same postulate as randomized evaluation, that even if the impact of a program varies across individuals, thinking of average treatment effects makes sense.

That being said, it is absolutely clear that each randomized evaluation is conducted in specific circumstances. Without a theory of why the program has the effect it has (which is typically hard to formulate, and harder to test), generalizing from one well-executed experiment may be unwarranted. In my view, this does not mean that we should give up on randomized evaluation, but that we should encourage replication of experiments in key domains of interests in several different settings. While there will always be the possibility that one program that was unsuccessful in one context would have been successful in another, it should be possible to, eventually, narrow these circumstances to such a narrow set that they become theoretical possibilities, rather than practical concerns. In the first section, we gave several examples of replication. Replication is

[6] A UNICEF representative once objected to the idea that randomized evaluations could be taught, and "were not nuclear physics". His answer was that "studying human being is much more complicated than nuclear physics."

one area where international organizations, which are present in most countries, can play a key role. It implies that they do not jump too quickly on the result of one experiment, but that they take the time to do more. An example of an opportunity which 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. This will be a fascinating set of studies.

Another dimension in which randomized evaluations are limited is that they typically evaluate only one specific variant of a program. While they can answer the question of the impact and the cost effectiveness of this specific variant, they have little to say about whether the variant that was tested was the optimal way to design the program. For example, was the specific level and structure of the PROGRESA program optimal? Or would smaller (or larger) transfers have been more cost effective? Some theoretical guidance will be needed to extrapolate from the results of a series of evaluations to obtain lessons that can be generalized. While this is clearly true, this applies to any kind of program evaluation. This suggests that one dimension of replication would be to try different variants of the program. Potentially, this does not even require a comparison group. After a few data points with different variants are obtained, with some theoretical guidance, it might be possible to interpolate between them. Obviously, not all combinations of possible programs in all contexts can be evaluated. Theory needs to guide the choice of the programs to be evaluated, and which variations are likely to matter. There will then be a useful interaction between theory and practice, with the evaluation of real programs serving as the test for the theory.

An alternative is to use the exogenous variation created by the randomization to help identify a structural model. Attanasio et al. (2001) and Berhman et al. (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.

2.3 The Political Economy of Program Evaluation

The objections to randomized evaluations thus far point to real limitations, but none of these limitations seem serious enough to justify their rarity. In this case, why are they so rare? Cook (2001) attributes their rarity in education to the post-modern culture in American education schools, hostile to the traditional conception of causation which underlies statistical implementation.

Pritchett (2002) proposes a political economy argument. He 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 impact to voters. Kremer (2003) proposed a complementary explanation: In this view, policy makers are not systematically fooled, but have difficulty gauging the quality of evidence. Advocates can suppress unfavorable evaluation results. Kremer writes:

> "Suppose retrospective regressions yield estimated program effects equal to the true effect, plus measurement error, plus a bias term, possibly with mean zero. Program advocates then select the highest estimates to present to policy makers, while any opponents select the most negative estimates. Knowing this, policy makers 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 there is little incentive to conduct randomized evaluations. Since the resulting estimates include no bias term, they are unlikely to be high enough or low enough that advocates will present them to policy makers. Even if results are presented to policy makers, policy makers unable to gauge the quality of particular studies will discount them. Why fund a project that a randomized evaluation suggests has a 25% rate of return when advocates of competing projects claim a 100% rate of return?" (Kremer, 2003, p. 10)

In this world, an international organization can play two key roles, by demanding randomized evaluation before funding a program, and by acting as a certification agency for evaluation. Moreover, if it becomes easier for policy makers 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.

3. Scaling Up and Evaluation in International Organizations

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

3.1 **Defining Priorities for Evaluation**

It is almost certainly counter-productive to demand that *all projects* be evaluated. Clearly, all projects need to be monitored to make sure that they actually happened. 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 projects 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.

3.2 Setting up Autonomous Impact Evaluation Units

Not all projects need an impact evaluation, but all projects will continue to be subject to regular monitoring and evaluation; data will be collected in the course of the project, and it is likely that the project advocates within the organizations will be tempted to use it to generate "estimates" of the impact of the program. If these evaluations are mixed up with credible impact evaluation, they will dilute their impact. Banerjee and He (2003) argue that the World Bank's decision and reports have little impact on market decisions as well as 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 justified their decisions. Credibility would require the Bank to "firewall" the "real" impact evaluation from the rest of the organization. A special unit should be set up, potentially within the research unit, to emphasize the distinction with regular monitoring and evaluation procedures. The results of studies produced or endorsed by the unit would be published separately from other World Bank documents.

For clarity, the unit should have a simple mandate: Encourage, conduct, and finance randomized impact evaluations, and disseminate the results. It should also encourage data collection and study of true "natural experiments" with program-induced randomization. As we mentioned in Section 1, 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 of 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 should have a large dollar budget, used to finance and conduct randomized evaluations of internal and external projects. The unit should have its own evaluation projects, in the key areas identified by the organization.

3.2 Work with Partners

The unit should also work 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 Banks 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.

3.3. Certify and Disseminate Evaluation Results

Another role the unit could serve, after establishing a reputation for quality, is that of a certifying body and "clearing house" and dissemination agency. In order to be useful, evaluation results need to be accessible to practitioners, within and outside development agencies. A key role of the unit should 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 policy makers, this knowledge is as useful as knowing 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.

4. Conclusion: Using Evaluation to Build Long Term Consensus for Development

International organizations have the opportunity to play a leading role in leveraging randomized evaluation to achieve real scaling up. By promoting, financing, and disseminating the results of impact evaluations, they can increase the effectiveness of development aid and provide information to policy makers and international donors. The impact can be further enhanced if this encourages other institutions (private and international donors) to demand and finance quality impact evaluation. Moreover, by establishing a clear standard of what works and what does not, they can build long term consensus for development, by increasing their credibility vis-à-vis their potential donors. This would be an international public good at the heart of the role of international organizations.

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 D., and Alan B. Krueger (1999) "Empirical strategies in labor economics," in Handbook of Labor Economics, Vol. 3A. Orley Ashenfelter and David Card (Eds.). Amsterdam: North Holland, pp. 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 D., 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.

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

Banerjee, Abhijit, Shawn Cole, Esther Duflo and Leigh Linden (2003) "Improving the Quality of Education in India: Evidence from Three Randomized Experiments," mimeo, MIT.

Banerjee, Abhijit and Ruimin He (2003) "The World Bank of the Future," forthcoming in *American Economic Review, Papers and Proceedings*.

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, 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 (2002) "How Much Should We Trust Difference in Differences Estimates?" NBER working paper #8841.

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

Campbell, Donald T. (1969) "Reforms as experiments," *American Psychologist*, 24, 407-429.

Card, David (1999) "The causal effect of education on earnings," in Handbook of Labor Economics, Vol. 3A. Orley Ashenfelter and David Card (Eds.). Amsterdam: North Holland, pp. 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.

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

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 (2000) "Retrospective vs. prospective analyses of school inputs: The case of flip charts in Kenya," NBER Working Paper #8018.

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," forthcoming in *American Economic Review Papers and Proceedings,* May 2003.

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

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), Empowerment and Poverty Reduction: A Sourcebook, The World Bank, 2000.

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: World Bank, 2000).

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), pp. 958-996.

Pritchett, Lant (2002) "It Pays to be Ignorant: A Simple Political Economy of Rigorous Program Evaluation," mimeo, JFK School of Government, Harvard University.

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*.