Use of Randomization in the Evaluation of Development Effectiveness[1]

Esther Duflo[2]   Michael Kremer[3]

Paper prepared for the World Bank Operations Evaluation Department (OED)
Conference on Evaluation and Development Effectiveness
in Washington, D.C.
15-16 July, 2003

Abstract

In this paper, we argue that current evaluation practices often encounter problems that prevent them from effectively determining program impact, and that there is considerable scope for greater use of randomized evaluation methods in addressing these problems. We review recent randomized evaluations of educational programs in developing countries, including programs to increase school participation, provide educational inputs, and reform education, and then extract lessons from these examples. Finally, we discuss the role the international agencies can play in promoting and financing rigorous impact evaluations, including randomized evaluations.

[2] Department of Economics, MIT, BREAD, and NBER, eduflo@mit.edu
[3] Department of Economics, Harvard University, BREAD, The Brookings Institution, Center for Global Development, and NBER, mkremer@fas.harvard.edu

Use of Randomization in the Evaluation of Development Effectiveness [4],[5]

Esther Duflo[6]   Michael Kremer[7]

Paper prepared for the World Bank Operations Evaluation Department (OED)
Conference on Evaluation and Development Effectiveness
in Washington, D.C.
15-16 July, 2003

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

The benefits of knowing which programs work and which do not extend far beyond any program or agency, and credible impact evaluations are global public goods in the sense that they can offer reliable guidance to international organizations, governments, donors, and NGOs beyond national borders.

For a broad class of development programs, randomized evaluations can be used to 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 uses the case of educational programs in developing countries as an example.

It is important to note that we are not proposing that all projects be subject to randomized evaluations. We are arguing that there is currently a tremendous imbalance in evaluation methodology, and that increasing the share of projects subject to randomized evaluation from near-zero to even a small fraction could have a tremendous impact. All too often development policy is based on fads, and randomized evaluations could allow it to be based on evidence.

This paper proceeds as follows: in Section 1 we discuss the methodology of randomized evaluations: we present the impact evaluation problem, review why other current evaluation methods may often be unable to adequately control for selection bias, and discuss why randomized evaluations can be useful in addressing the problems encountered by other evaluation practices. In Section 2 we review recent randomized evaluations of educational programs in developing countries, including programs to increase school participation, provide educational inputs, and reform education. In Section 3 we extract lessons from the evaluations described in Section 2. In Section 4 we review an example of current practice, offer political economy explanations for why randomized evaluations are so rare, and discuss the role the international agencies can play in promoting and financing rigorous evaluations (including randomized evaluations). Section 5 concludes by discussing the value of credible impact evaluations as international public goods.

1. <u>The methodology of randomized evaluations</u>

In Subsection 1.1 we discuss the selection bias problem that can arise when conducting impact evaluations, and in Subsection 1.2 we discuss non-randomized evaluation methods that are used in attempting to control for this bias.

1.1 The evaluation problem

Any impact evaluation attempts to answer an essentially counterfactual question: how would individuals who participated in the program have fared in the absence of the program? How would those who were not exposed to the program have fared in the presence of the program? The difficulty with these questions is immediate: at a given

point in time, an individual is observed to be either exposed or not exposed to the program. Comparing the same individual over time will not, in most cases, give a reliable estimate of the impact the program had on him or her, since many other things may have changed at the same time as the program was introduced. We 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 of individuals who were not exposed to the program. The critical objective of impact evaluation is therefore to establish a credible comparison group, a group of individuals who *in the absence of the program* would have had outcomes similar to those who were exposed to the program. This group should give us an idea of what would have happened to the members of the program group if they had not been exposed, and thus allow us to obtain an estimate of the average impact on the group in question.

However, in reality it is generally the case that individuals who participated in a program and those who did not are 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, in addition, 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, and any differences between the groups can 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 typically cannot decompose the overall difference into a treatment effect and a bias term.

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

As we will see in Section 2, the random selection of treatment and comparison groups can occur in several circumstances: Subsection 2.1.1 discusses (using the example of PROGRESA, a program designed to increase school participation in Mexico) how prospective randomized evaluations can be used and how their results can help in scaling successful programs; Subsection 2.1.2 discusses (using examples of school-based health programs in Kenya and India) how prospective randomized evaluations can be used when implementing adapted replications of programs; and Subsection 2.3.2 discusses (using the example of a school voucher program in Colombia) how program-induced randomization can occur.

Natural or organized randomized evaluations are not the only methodologies that can be used to obtain credible impact evaluations of program effects. Researchers have developed alternative techniques to control for bias as well as possible, and progress has been made, most notably by labor economists (there are numerous excellent technical and non-technical surveys of these techniques as well as their value and limitations; see Angrist and Krueger, 1999 and 2001; Card, 1999; and Meyer, 1995). In Subsection 1.2, we will briefly review some of the techniques that are most popular with researchers: propensity score matching, difference-in-difference estimates, and regression discontinuity design. However, as we will discuss at the end of Subsection 1.2, the main issue that arises with these techniques is that they may contain large and unknown biases resulting from specification errors.

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

a large scale, the programs could affect the functioning of the school system and thus have a different impact.

1.2 Other techniques to control for selection and other omitted variable bias

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

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

laws and found that with about twenty years of data, difference-in-difference estimates found an "effect" significant at the 5% level of up to 45% of the placebo laws. As an example of where difference-in-difference estimates can be used, Duflo (2001) took advantage of a rapid school expansion program that occurred in Indonesia in the 1970s to estimate the impact of building schools on schooling and subsequent wages. Identification 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 participating in the program are easily identified (children twelve years or older when the program started did not participate in the program). The increased growth of education across cohorts in regions that received more schools suggests that access to schools contributed to increased education. The trends were quite parallel before the program and shifted clearly for the first cohort that was exposed to the program, thus reinforcing confidence in the identification assumption. However, this identification strategy is not usually valid; often when policy changes are used to identify the effect of a particular policy, the policy change is itself endogenous to the outcomes it was meant to affect, thus making identification impossible (see Besley and Case, 2000).

Finally, a third strategy, called "regression discontinuity design" (see Campbell, 1969) takes advantage of the fact that program rules sometimes generate discontinuities that can be used to identify the effect of the program by comparing those who made it to those who "almost made it." That is, if resources are allocated on the basis of a certain number of points, it is possible to compare those just above to those just below the threshold. Angrist and Lavy (1999) use this technique to evaluate the impact of class size in Israel. In Israel, a second teacher is allocated every time the class size grows above 40. This policy generates discontinuities in class size when the enrollment in a grade grows from 40 to 41 (as class size changes from one class of 40 to one class each of size 20 and 21), 80 to 81, etc. Angrist and Lavy compared test scores in classes just above and just below this threshold, and found that those just above the threshold have significantly higher test 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.[8] Such discontinuities in program rules, when enforced, are thus sources of identification. However, such discontinuities are often *not* implemented, especially in developing countries. For example, researchers attempted to use as a source of identification the discontinuity in a Grameen bank (the flagship microcredit organization in Bangladesh) policy, 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.

As mentioned in Subsection 1.1, the main issue that arises with these techniques is that they may contain large and unknown biases resulting from specification errors. LaLonde (1986) found that many of the econometric procedures and comparison groups used in program evaluations did not yield accurate or precise estimates, and that such econometric estimates often differ significantly from experimental results.

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

2. **Examples of randomized evaluations of educational programs**

---

[8] Angrist and Lavy note that parents who discover they received a bad draw in the "enrollment lottery" (*e.g.*, an enrollment of 38) might then elect to pull their kids out of the public school system and into private schools. However, as Angrist and Lavy discuss, private elementary schooling is rare in Israel outside of the ultra-orthodox community.

In this section, we present recent randomized evaluations of educational programs in developing countries: Subsection 2.1 discusses programs designed to increase school participation, Subsection 2.2 discusses programs to provide educational inputs, and Subsection 2.3 discusses educational reform programs. In Section 3 we will extract lessons from these examples.

2.1 Increasing school participation

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

In this subsection, we review recent randomized evaluations of programs designed to increase school participation in developing countries. As one obvious means of increasing school participation is to decrease or remove financial barriers, in this subsection we review recent randomized evaluations of programs designed to increase school participation in developing countries through reducing cost of school, or even paying for school attendance: Subsection 2.1.1 discusses the PROGRESA program, and Subsection 2.1.2 discusses research on various programs which decrease costs of education and then compares the cost-effectiveness of several different interventions.[9]

2.1.1   PROGRESA

Carefully constructed program evaluations form a sound basis for decisions of whether or not to scale up existing projects, as positive results can help to build a

consensus for a project. The PROGRESA program in Mexico, designed to increase school participation, is a striking example of this phenomena. PROGRESA provides cash 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). When the program was launched in 1998, officials in the Mexican government made a conscious decision to take advantage of the fact that budgetary constraints made it impossible to reach the 50,000 potential participant communities of PROGRESA immediately, and instead began with a program in 506 communities. Half of those communities were randomly selected to receive the program, and baseline and subsequent data were collected in the remaining communities (Gertler and Boyce, 2001). Part of the rationale for this decision was to increase the probability that the program would be continued in case of a change in the party in power, as the proponents of the program understood that the program would require continuous political support in order to be scaled up successfully. The task of evaluating the program was given to academic researchers through the International Food Policy Research Institute (IFPRI), the data was made accessible to numerous researchers, and a number of papers have been written on PROGRESA's impact (most of which are accessible on the IFPRI web site). The evaluations show that the program was effective in improving both health and education: comparing PROGRESA participants and non-participants, Gertler and Boyce (2001) show that children on average had 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 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 randomized phase-in of the program allowed such clear documentation of the program's positive effects, PROGRESA was indeed maintained when the Mexican government changed hands: by 2000, PROGRESA was reaching 2.6 million families (10% of the families in Mexico) and had a budget of US $800 million, or 0.2% of GDP (Gertler and Boyce, 2001). The program was subsequently expanded to urban

---

[9] By school participation, we mean that we will focus on a comprehensive measure of school participation: a pupil is considered a participant if she or he is present in school on a given day, and a non-participant if

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 both by the expansion of the program as well as by the tremendous visibility that the program acquired.

2.1.2   School meals, cost of education, and school health in Kenya: comparing the cost-effectiveness of different interventions

A central and fundamental policy concern for developing countries is the relative cost-effectiveness of various interventions intended to increase school participation. Evaluations of cost-effectiveness require knowledge of a program's costs as well as its impact, and comparability across studies requires some common environment. It is difficult to compare the impact of PROGRESA's cash transfers with, say, school meals in Kenya, since it is unclear whether the resulting differences are associated with the type of program or the larger environment. In general, analysts and policy makers are left with a choice between retrospective studies, which allow comparison of different factors affecting school participation, and randomized evaluations, which yield very credible estimates of the effect of single programs. One exception to our general inability to compare cost-effectiveness estimates is the recent set of studies of programs seeking to improve school participation conducted in Kenya: by evaluating a number of programs in a similar setting (a specific district in Western Kenya), it is possible to explicitly compare the cost-effectiveness of different approaches to increasing school participation.

Vermeersch examines the effect of school meals on school participation in Kenya. School participation was 30% greater in twenty-five Kenyan pre-schools where a free breakfast was introduced than in twenty-five comparison schools. However, the provision of meals did cut into instruction time. Overall, test scores were .4 standard deviations greater in the program schools, but only if the teacher was well trained prior to the program (Vermeersch, 2002).

 Kremer *et al.* (2002) evaluate a program in which a non-governmental organization (NGO), Internationaal Christelijk Steunfonds Africa (ICS), provided

---

she or he is not in school on that day.

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

Poor health may also limit school participation: for example, intestinal helminthes (such as hookworm) affect a quarter of the world's population, and are particularly prevalent among school-age children. Miguel and Kremer (2003a) evaluate a program of twice-yearly school-based mass treatment with inexpensive de-worming 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 seven percentage points) lower than in comparison schools. Including the spillover effect, the program increased schooling by 0.15 years per person treated.

Because these programs were conducted in similar environments, cost-effectiveness estimates from numerous randomized evaluations can be readily compared. Deworming was found to be extraordinarily cost-effective at only $3.50 per additional year of schooling (Miguel and Kremer, 2003a). In contrast, even under optimistic assumptions provision of free uniforms would cost $99 per additional year of school

participation induced (Kremer *et al.,* 2002). The school meals program, which targeted pre-schoolers rather than primary school age children, cost $36 per additional year of schooling induced (Vermeersch, 2003). This suggests that school health programs may be one of the most cost-effective ways of increasing school participation.

2.2 School inputs

In this subsection, we review recent randomized evaluations of programs which provide various inputs to schools in developing countries: Subsection 2.2.1 compares the results of retrospective and prospective research on the provision of textbooks and flipcharts in Kenya, Subsection 2.2.2 discusses a program which placed additional teachers in non-formal education centers in India, and Subsection 2.2.3 reviews the evaluation of a remedial education program in India.

2.2.1 Retrospective and prospective studies of inputs in Kenyan primary schools

Many are skeptical about the effects of educational inputs on learning based on existing retrospective evaluations (Hanushek, 1995). One potential weakness of such evaluations is that observed inputs may be correlated with omitted variables that affect educational outcomes. For example, this bias could be upward if observed inputs are correlated with unobserved parental or community support for education, or this bias could be downward if compensatory programs provide assistance to poorly performing schools.

Although retrospective studies provide at best mixed evidence on the effect of many types of school inputs, they typically suggest that provision of additional textbooks in schools with low initial stocks can improve learning. Indeed, cross-sectional and difference-in-difference analyses of Kenyan data would suggest that textbooks have dramatic effects on test scores. Results from a randomized evaluation, however, point to a subtler picture. Provision of textbooks increased test scores by about 0.2 standard deviations, but only among students who had scored in the top one or two quintiles on pre-tests prior to the program. Textbook provision did not affect scores for the bottom 60% of students (Glewwe *et al.*, 2002). Many students may have failed to benefit from textbooks because they had difficulty understanding them: Kenyan textbooks are in

English, the official language of instruction, but English is most pupils' third language (after their mother tongue and Swahili). More generally, the Kenyan curriculum is set at a level that, while perhaps appropriate for elite families in Nairobi, is far ahead of the level typically attained by rural students given high levels of student and teacher absence.

Given the results of the textbook study, researchers tried providing flipcharts, an alternative input that presumably was more accessible to weak pupils. Glewwe *et al*. (forthcoming) 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.

2.2.2 Placing additional teachers in non-formal education centers

Banerjee *et al.* (2000) evaluated a program in which Seva Mandir, an Indian NGO, placed a second teacher in non-formal education centers the NGO runs in Indian 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 (when possible, a woman) was randomly assigned to 21 out of 42 schools, and the hope was to increase the number of days the

school was open, increase children's participation, and increase performance by providing more individualized attention to the children. By providing a female teacher, the NGO also hoped to make school more attractive for girls. Teacher attendance and child attendance were regularly monitored 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. Girls' attendance increased by 50%. However, there were no differences in test scores. It is worth noting that careful evaluations form a sound basis for decisions of whether or not to scale up existing projects. 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, and the savings were then used to expand other programs.

2.2.3 Remedial education programs

Pratham, an Indian NGO, implemented a remedial education program in 1994 that now reaches over 161,000 children in twenty cities. The program hires young women from the communities to provide remedial education in government schools to children who have reached grade 2, 3 or 4 without having mastered the basic grade 1 competencies. Children who are identified as lagging 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 the NGO's flagship interventions, at the same time as they were looking to expand the program; the expansion into a new city, Vadodara, provided an opportunity to conduct a randomized evaluation (Banerjee, Cole, Duflo and Linden, 2003). In the first year (1999-2000), the program was expanded to forty-nine (randomly selected) of the 123 Vadodara government schools. In 2000-2001, the program was expanded to all the schools, but the half the schools received a remedial teacher for grade 3, and half received one for grade 4. Grade 3 students in schools that were exposed to the program in grade 4 serve as the comparison group for grade 3 students who were directly exposed to the program. Simultaneously, a similar intervention was conducted in a district of Mumbai, where half the schools received the remedial teachers in grade 2, and half received teachers in grade 3.   The program was continued for an additional year, with the school switching groups. The program was thus conducted in several grades, in

two cities, and with no school feeling that they were deprived of resources relative to others, since all schools participated in the program. On average, after two years the program increased student test scores by 0.39 standard deviations. Moreover, the gains are largest for children at the bottom of the distribution: children in the bottom third gained 0.6 standard deviations after two years. The impact of the program is rising over time, and is very similar across cities and child gender. Hiring remedial education teachers from the community appears to be ten times more cost-effective than hiring new teachers. One can be relatively confident in recommending the scaling up of this program, at least in India, on the basis of these estimates since the program was continued for a period of time, was evaluated in two very different contexts, and has shown its ability to be rolled out on a large scale.

2.3 School reform

There is reason to believe that many school systems could benefit from considerable reform. For example, evidence from Kenyan evaluations discussed previously suggests that budgets are misallocated and that the curriculum focuses excessively on the strongest students. Moreover, teacher incentives in Kenya, as in much of the developing world, are quite weak. Absence among teachers is quite high, around 20%. A number of school reform initiatives have been proposed, ranging from decentralization of budget authority to strengthening links between teacher pay and performance to vouchers and school choice. As an example, a decentralization program in Kenya that provided small grants to parent-run school committees induced them to purchase textbooks, with educational consequences similar to those of the textbook program (Glewwe, Ilias, and Kremer, 2003). Providing larger grants led school committees to shift spending toward construction, and no educational impact could be observed from this, at least in the short-run. In Subsection 2.3.1 we discuss a teacher incentive program that arose from this decentralization program in Kenya, and Subsection 2.3.2 reviews the results of a school voucher program in Colombia.

2.3.1   Teacher incentives

Some parent-run school committees in Kenya provide gifts to teachers whose students perform well. Glewwe *et al.* (2003) evaluate a program that provided prizes to teachers in schools that performed well on exams and had low dropout rates. In theory, this type of incentive could lead teachers to either increase effort or, alternatively, to teach to the test. Empirically, teachers responded to the program not by increasing attendance but rather by increasing prep sessions designed to prepare students for the exams. Consistent with a model in which teachers responded to the program primarily by increasing effort devoted to manipulating test scores rather than by increasing effort at stimulating long-term learning, text scores for pupils who had been part of the program initially increased but then fell back to levels similar to the comparison group at the end of the program.

2.3.2   School vouchers

Angrist *et al*. (2002) evaluate a Colombian program in which vouchers for private schools were allocated by lottery because of limitations in the program's budget. Vouchers were renewable conditional on satisfactory academic performance. The researchers found that lottery winners were 15-20% 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. The effects of the program were larger for girls than for boys. 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.

3. Lessons

The evaluations described in Section 2 offer both substantive and methodological lessons. School participation can be substantially increased through implementing inexpensive health programs, reducing the costs of school to households, or providing school meals. Given the existing education system in Kenya, which like many developing countries has a curriculum focused on the strongest students, limited teacher

incentives, and sub-optimal budget allocation, simply providing more resources may have a limited impact on school quality. A remedial education program in India suggests that it is possible to improve student test scores substantially at a very low cost. Decentralizing budgets to school committees or providing teacher incentives based on test scores had little impact in Kenya, but a school choice program in Colombia yielded dramatic benefits for participants. In this section, we review some of the methodological lessons that can be drawn from the examples discussed in Section 2.

3.1 Results from randomized evaluations can be quite different from those drawn from retrospective evaluations

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

Comparative studies which estimate a program's impact using experimental methods and then re-estimate impact using one or several different non-experimental methods suggest that omitted variable bias is a significant problem beyond just the examples mentioned here. Although we are not aware of any systematic review of studies in developing countries, one recent study in developed countries suggest that omitted variable bias is a major problem when non-experimental methods are used (Glazerman, Levy, and Meyers, 2002). This studies assessed both experimental and non-experimental methods in the context of welfare, job training, and employment service programs and found that non-experimental estimators often produce results dramatically different from randomized evaluations, that the estimated bias is often large, and no strategies seems to perform consistently well.[10]

---

[10] One recent study not included in the analysis of Glazerman, Levy, and Meyers (2002) is that of Buddlemeyer and Skoufias (2003). Buddlemeyer and Skoufias use randomized evaluation results as a benchmark to examine the performance of regression discontinuity design for evaluating the impact of the

Future research along these lines would be valuable, as such comparative studies can be useful in assessing the extent to which the biases of retrospective estimates are significant. However, when the comparison group for the non-experimental portions of these comparative studies is decided *ex post*, the evaluator may be able to pick from a variety of plausible comparison groups, some of which may have results which match experimental estimates and some of which may not (as discussed in Subsection 3.6, this is also an issue for retrospective studies in regard to problems with publication bias). Possible ways of addressing these concerns in the future include conducting non-experimental evaluations first (before the results of randomized evaluations are released), or having researchers conduct blind non-experimental evaluations without knowledge of the results of randomized evaluations or other non-experimental studies.

3.2 Randomized evaluations are often feasible

As is clear from the examples discussed in this paper, randomized evaluations are feasible and have been conducted successfully. Randomized evaluations are labor-intensive and costly, but no more so than other data collection activities.

Political economy concerns over randomized evaluations (as we will discuss in Subsection 4.2) may sometimes make it difficult to not implement a program in the entire population: for example, "opportunidades," the urban version of PROGRESA, will not start with a randomized evaluation because of the strong opposition to delaying some people access to the program. However, these concerns can be tackled at several levels: as we will discuss in Subsection 3.3, financial constraints often necessitate phasing-in programs over time, and randomization may actually be the fairest way of determining the order of phase-in.

3.3 NGOs are well-suited to conduct randomized evaluations, but will require technical assistance (for example, from academics) and outside financing

Governments are far from being the only possible outlets through which randomized evaluations can be organized. Indeed, the evidence presented in this paper

---

PROGRESA program on child health and school attendance and found the performance of regression discontinuity design in this case to be good.

suggests one possible model is that of evaluation of NGO projects. Unlike governments, NGOs are not expected to serve entire populations. Financial and administrative constraints often lead NGOs to phase in programs over time, and randomization will often be the fairest way of determining the order of phase-in. Even small NGOs can substantially affect budgets in developing countries. Given that many NGOs exist and that they frequently seek out new projects, it is often relatively straightforward to find NGOs willing to conduct randomized evaluations: hitches are more often logistical than philosophical. For example, the set of recent studies conducted in Kenya have been carried out through a collaboration with the Kenyan NGO Internationaal Christelijk Steunfonds (ICS) Africa: ICS was keenly interested in using randomized evaluations to see the impact its programs are having, as well in sharing credible evaluation results with other stake holders and policy makers. A second example is the collaboration between the Indian NGO Pratham and MIT researchers which led to the evaluations of the remedial education and computer-assisted learning programs (Banerjee *et al.*, 2003). This collaboration was initiated when Pratham was looking for partnership to evaluate their programs; Pratham understood the value of randomization and was able to convey the importance of such evaluations to the schoolteachers involved in the project. However, while NGOs are well placed to conduct randomized evaluations, it is less reasonable to expect them to finance these evaluations. The evaluations of the ICS deworming programs were made possible by financial support from the World Bank, the Partnership for Child Development, and the US National Institutes of Health (NIH), and the MacArthur Foundation. In the case of the Indian educational programs, Pratham was able to find a corporate sponsor: India's second-largest bank, ICICI Bank, was keenly interested in evaluating the impact of the program and helped to finance part of the evaluation. In general, given that accurate estimates of program effects are (as we will discuss in Section 5) international public goods, we argue that randomized evaluations should be financed internationally. As we will discuss in Section 4, international organizations could finance randomized evaluations organized in collaboration between researchers (from those organizations as well as from academia) and NGOs, and the potential is enormous.

3.4 Costs can be reduced and comparability enhanced by conducting a series of evaluations in the same area

Once staffs are trained, they can work on multiple projects. Since data collection is the most costly element of these evaluations, crosscutting the sample can also dramatically reduce costs. For example, many of the various programs seeking to increase school participation where implemented in the same area, by the same organization. The teacher incentives (Glewwe, Ilias, and Kremer, 2003) and textbook (Kremer *et al.*, 2002) programs were evaluated in the same 100 schools: one group had textbooks only, one had textbooks and incentives, one had incentives only, and one had neither. The effect of the incentive program should thus be interpreted as the effect of an incentive program conditional on half the schools having extra textbooks. Likewise, a computer assisted learning program was implemented in Vadodara in the same set of schools where the remedial education study was conducted. This tactic must take into account potential interactions between programs (which can be estimated if the sample is large enough), and may not be appropriate if one program makes the schools atypical.

3.5 Randomized evaluations have a number of limitations, but many of these limitations also apply to other techniques

Many of the limitations of randomized evaluations also apply to other techniques. In this subsection we will review four issues which affect both randomized and non-randomized evaluations (sample selection bias, attrition bias, behavioral responses, and spillover effects), and we will argue that randomized methods often allow for easier correction for these limitations than do non-randomized methods.

Sample selection problems could arise if factors other than random assignment influence program allocation. For example, parents may move their children from a class (or a school) without the program to a school with the program. Conversely, individuals allocated to a treatment group may not receive the treatment (for example, because they decide not to take up the program). Even if randomized methods have been employed and the intended allocation of the program was random, the actual allocation may not be. This problem can be addressed through "intention to treat (ITT)" methods or by using

random assignment as an instrument of variables for actual assignment. 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 and scale up the difference by dividing it by the difference in the probability of receiving the treatment in those two groups to obtain the local average treatment effect estimate (Imbens and Angrist, 1994). Methods such as ITT estimates allow selection problems to be addressed fairly easily in the context of randomized evaluations, but it is often much more difficult to make these corrections in the case of a retrospective analysis.

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

Finally, programs may create spillover effects on people who have themselves not been treated. These spillovers may be physical, as found for the Kenyan de-worming program by Miguel and Kremer (2003a) when de-worming interferes with disease transmission and thus makes children in treatment schools (and in schools near treatment schools) less likely to have worms even if they were not themselves given the medicine.

Such spillovers might also operate through prices, as found by Vermeersch (2002) when the provision of school meals leads competing local schools to reduce school fees. Finally, there might also be learning and imitation effects (Duflo and Saez, forthcoming; Miguel and Kremer, 2003b). If such spillovers are global (for example, due to changes in world prices), identification of total program impacts will be problematic with any methodology. However, if such spillovers are local then randomization at the level of groups can allow estimation of the total program effect within groups and can generate sufficient variation in local treatment density to measure spillovers across groups. For example, the solution in the case of the de-worming study was to choose the *school* (rather than the pupils within a school) as the unit of randomization (Miguel and Kremer, 2003a), and to look at the number of treatment and comparison schools within neighborhoods (of course, this requires a larger sample size).

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

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

3.6 Publication bias appears to be substantial with retrospective studies; randomized evaluations can help address publication bias problems, but institutions are also needed

Publication bias is a particularly important issue that must be addressed: available evidence suggests the publication bias problem is severe (DeLong and Lang, 1992) and, in particular, the problem of publication bias appears to be significant with studies employing non-experimental methods. 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 are much more interested and able to publish

positive results than modest or insignificant results. However, clearly many programs fail, and publication bias will be substantial if positive results are much more likely to be published.

Publication bias is likely to a particular problem with retrospective studies. *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. 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 these estimates 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 the 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 the results are usually documented and published even if the results suggest quite modest effects or even no effects at all.

As we will discuss in Subsection 4.3 it is important to put institutions in place to ensure negative results are disseminated. Such a system is already in place for medical trial results, and creating a similar system for documenting evaluations of social programs would help to alleviate the problem of publication bias. Beyond allowing for a clearer picture of which interventions have worked and which have not, this type of institution would provide the level of transparency necessary for systematic literature reviews to be less biased in their conclusions about the efficacy of particular policies and programs.

3.7 Although any given randomized evaluation is conducted within a specific framework with unique circumstances, randomized evaluations can shed light on general issues

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

Often the results of the first phase of a project may be difficult to interpret because of circumstances unique to it being the first phase: if the project is unsuccessful, it may be because it faced implementation problems that could be avoided in later phases of the project; if the project is successful, it may be because more resources were allocated to it than would have been under a more realistic situation or in a less favorable context. Even if the choice of the comparison and treatment groups ensures internal validity of estimates, any method of evaluation is subject to problems with external validity due to the specific circumstances of implementation (that is, the results may not be able to be generalized to other contexts). Specific to randomized evaluations are problems such as members of the treatment group changing their behavior (known as the Hawthorne effect) and members of the comparison group having their behavior affected (known as the John Henry effect) as a result of being part of the randomized evaluation.

A first answer to some of these concerns is to implement adapted replications of successful (and potentially unsuccessful) programs in different contexts. Such adapted replications present two advantages: first, in the process of "transplanting" a program, circumstances will change and robust programs will show their effectiveness by surviving these changes; second, obtaining several estimates in different contexts will provide some guidance about whether the impacts of the program are notably different in different groups. Replication of the initial phase of a study in a 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 participants be phased into the program in random order. In addition, such adapted replications can be used to check whether program effects within samples vary with covariance: for example, suppose that the effect of a given program is smaller in schools with good teachers. One might consider whether in a different setting with much better teachers the effect would be smaller.

One example is the work in India of Bobonis, Miguel, and Sharma (2002), which conducted an adapted replication of the deworming study in Kenya. 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; thus, the program was modified to include iron supplementation. The program was administered through a network of pre-schools in urban India. After one year of treatment, the researchers 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). This supports the conclusion of the deworming research in Kenya (Miguel and Kremer, 2003a) that school health programs may be one of the most cost-effective ways of increasing school participation and, importantly, suggest that this conclusion may be relevant in low-income countries outside of Africa.

It is worth noting that the exogenous variation created by randomization can be used to help identify a structural model. Attanasio *et al*. (2001) and Berhman *et al*. (2002) are two examples of using this exercise in combination with the PROGRESA data

to make some prediction of the possible effect of varying the schedule of transfers. For example, Attanasio *et al.* (2001) found that the randomized component of the PROGRESA data induced extremely useful exogenous variation that helped in the identification of a richer and more flexible structural model. These studies rest on assumptions that one is free to believe or not, but at least these studies are freed of *some* assumptions by the presence of this exogenous variation. The more general point is that randomized evaluations do not preclude the use of theory or assumptions: in fact, they generate data and variation that can be useful in identifying some aspects of these theories. For example, evaluations suggest that the Kenyan educational system is heavily geared towards top students and that reallocating budgets within primary education could lead to considerably better outcomes, pointing to perverse incentives created by Kenya's mix of local and national school finance (see Kremer *et al.*, 2002; Glewwe *et al.*, 2002).

4. <u>The role international agencies can play</u>

In Subsection 4.1 we review an example of current practice that failed to provide opportunities for rigorous evaluations due to a lack of planning. Subsection 4.2 presents some political economy arguments for why randomized evaluations are so rare, and Subsection 4.3 goes on to discuss how the international agencies can support the use of credible evaluation methods, including randomized evaluations.

4.1 The District Primary Education Program (DPEP): an example of lost opportunity

The District Primary Education Program (DPEP) in India, the largest World Bank sponsored education program, is an example of a large program with potentially very interesting evaluations whose potential on this count was jeopardized by lack of planning. DPEP was meant to be a 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 involving teacher training, inputs, and classrooms that seeks to improve the performance of public education. 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 (as measured by low female literacy rates), but high *potential for improvement.* In particular, the first districts chosen to receive the program were selected "on the basis of their ability to show success in a reasonable time frame" (Pandey, 2000, 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 (difference-in-difference) 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 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 existing data, such as census or National Sample Survey (NSS) data.

4.2 Why then are randomized evaluations so rare? Some political economy arguments

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. So 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 (forthcoming) 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.

A complementary explanation is that policy makers are not systematically fooled, but rather have difficulty gauging the quality of evidence in part because advocates can suppress unfavorable evaluation results. 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 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, those 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?

4.3 Evaluation in international organizations

The discussion in the preceding sections naturally suggests what international organizations could do to support the use of credible evaluation methods, including randomized evaluations. This section discusses several roles the international agencies can play in promoting and financing rigorous evaluations.

It is almost certainly counter-productive to demand that *all projects* be subject to impact evaluations. Clearly, all projects need to be monitored to make sure that they actually happen and to avoid misuse of funds. However, some programs, such as central bank independence, simply cannot be evaluated with the methods discussed in this paper. Even among projects that could potentially be evaluated, not all projects need impact evaluations. In fact, the value of a poorly identified impact evaluation is very low and its cost, in terms of credibility, is high, especially if international organizations take a leading role in promoting quality evaluation. A first objective is thus to cut down on the number of wasteful evaluations. Any proposed impact evaluation should be reviewed by a committee before any money is spent on data collection in order to avoid wasting

money. The committee's responsibility would be to assess the ability of the evaluation to deliver reliable causal estimates of the project's impact. A second objective would be to conduct credible evaluations in key areas. In consultation with a body of researchers and practitioners, each organization should determine key areas where they will promote impact evaluations. Randomized evaluations could also be set up in other areas when the opportunity occurs.

Credible impact evaluations require a great deal of work and, in addition, the benefits of credible impact evaluations (as we will discuss in Section 5) extend far beyond those conducting the evaluation; these factors mean that incentives to conduct rigorous evaluations are less than socially optimal. One promising remedy is to embed within the institutional framework of international agencies structures which will provide sufficient incentives for evaluators. Given the current scarcity of randomized evaluations within the institutional environment of international organizations, there may be scope for setting up a specialized unit to encourage, conduct, and finance rigorous impact evaluations, and to disseminate the results. As we will briefly discuss in this section, the potential for such a unit is tremendous: there exists a ready-made potential supply of evaluators both within the international agencies themselves as well as within academia and collaborations with NGOs offer many opportunities for evaluating policies of wide relevance.

Such an evaluation unit would encourage data collection and the study of true "natural randomized evaluations" with program-induced randomization. As we mentioned in Section 2, randomized evaluations are not the only method of conducting good impact evaluations. However, such other evaluations are conducted much more routinely, while randomized evaluations are conducted much too rarely in light of their value and the opportunities to conduct them. Part of the problem is that no one considers conducting such evaluations to be their job, and hence no one invests sufficiently to conduct them. In addition, all evaluations have common features, and thus would benefit from a specialized unit with specific expertise. Since impact evaluations generate international public goods, the unit should have a budget that would be used to finance and conduct rigorous evaluations of internal and external projects. The unit should have its own evaluation projects in the key areas identified by the organization.

As previously discussed, the unit should also work with partners, especially 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 organization or from outsiders, and from there choose projects to support. The unit could also encourage replication of important evaluations by sending out calls for specific proposals. The project could then be conducted in partnership with people from the unit or other researchers (academics, in particular). The unit could provide both financial and technical support for the project, with dedicated staff and researchers. Over time, on the basis of the acquired experience, the unit could also serve as a more general resource center by developing and diffusing training modules, tools, and guidelines (survey and testing instruments, as well as software that can be used for data entry and to facilitate randomization – similar in spirit to tools produced by other units in the World Bank) for randomized evaluation. The unit could also sponsor training sessions for practitioners.

Another role the unit could serve, after establishing a reputation for quality, is that of a dissemination agency (a "clearing house" of some sort). In order to be useful, evaluation results must be accessible to practitioners both within and outside development agencies. A key 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 would ideally include all information that could be useful in interpreting the results (estimates, sample size, region and time, type of project, cost, cost-benefit analysis, caveats, etc.), as well as references to other related studies. The database could include both randomized and non-randomized impact evaluations satisfying some criteria, provided that the different types of evaluation are clearly labeled. Evaluations would need to satisfy minimum reporting requirements to be included in the database, and all projects supported by the unit would have to be included in the database, whatever their results.

As previously discussed, such a database would help alleviate the "publication bias" (or "drawer") problem, which may be substantial if only positive results are published.   Academic journals may not be interested in publishing results of failed programs, but from the policy makers' point of view knowledge about negative results is

just as useful as knowledge about successful projects. Comparable requirements are placed on all federally funded medical projects in the United States. Ideally, over time, the database would become a basic reference for organizations and governments, especially as they seek funding for their projects. This database could kick-start a virtuous circle, with donors demanding credible evaluations before funding or continuing projects, more evaluations being conducted, and the general quality of evaluation work rising.

5. Conclusion

Rigorous and systemic evaluations have the potential to leverage the impact of international organizations well beyond simply their ability to finance programs. Credible impact evaluations are international public goods: the benefits of knowing that a program works or does not work extend well beyond the organization or the country implementing the program.[11] Programs that have been shown to be successful can be adapted for use in other countries and scaled up within countries, while unsuccessful programs can be abandoned. Through promoting, encouraging, and financing rigorous evaluations (such as credible randomized evaluations) of the programs they support, as well as of programs supported by others, the international organizations can provide guidance to the international organizations themselves, as well as other donors, governments, and NGOs in the ongoing search for successful programs. Moreover, by credibly establishing which programs work and which do not, the international agencies can counteract skepticism about the possibility of spending aid effectively and build long-term support for development. Just as randomized trials for pharmaceuticals revolutionized medicine in the 20th Century, randomized evaluations have the potential to revolutionize social policy during the 21st.

---

[11] In fact, the benefits of a credible evaluation are often negative for the person or organization promoting the program.

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 Handbook of Labor Economics, Vol. 3A. Orley Ashenfelter and David Card (Eds.). 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 (2000) "A Review of Estimates of Schooling/Earnings Relationship, with Tests for Publication Bias," NBER working paper #7457.

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," *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, MIT.

Banerjee, Abhijit and Michael Kremer with Jenny Lanjouw and Peter Lanjouw (2002) "Teacher-Student Ratios and School Performance in Udaipur, India: A Prospective Evaluation," mimeo, Harvard University.

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.

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 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.

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) (December), 1257-72.

Duflo, Esther (forthcoming) "Scaling Up and Evaluation," *Annual Bank Conference*

in Development Economics Conference Proceedings, The World Bank.

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

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

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

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

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, *Journal of Development Economics.*

Glewwe, Paul, Michael Kremer, and Sylvie Moulin (2002) "Textbooks and Test Scores: Evidence from a Prospective Evaluation in Kenya," mimeo, Harvard University.

Hanushek, Eric A. (1995) "Interpreting Recent Research on Schooling in Developing Countries," World Bank Research Observer, 10 (August), 227-246.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd (1998) "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66(5), 1017-98.

Heckman, James, Hidehiko Ichimura, and Petra Todd (1997) "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program," *Review of Economic Studies*, 64(4), 605-54.

Heckman, James, Robert Lalonde, and Jeffrey Smith (1999) "The Economics and Econometrics of Active Labor Market Programs," in *Handbook of Labor Economics: Volume 3*. O. Ashenfelter and D. Card, eds. North Holland.

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

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.

Kremer, Michael, Sylvie Moulin, and Robert Namunyu (2002) "Unbalanced Decentralization," mimeo, Harvard University.

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," *The 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 (2003a) "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," forthcoming in *Econometrica*.

Miguel, Edward and Michael Kremer (2003b) "Social Networks and Learning About Health in Kenya," mimeo, Harvard University.

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 (2003) "It Pays to be Ignorant," forthcoming, *Journal of Policy Reform.*

Rosenbaum, Paul R. (1995) "Observational Studies," In Series in Statistics. New York; Heidelberg; London: Springer.

Sen, Amartya (2002) "The Pratichi Report." Pratichi India Trust.

Shultz, T. Paul (2003) "School Subsidies for the Poor: Evaluating the Mexican PROGRESA Poverty Program," forthcoming, *Journal of Development Economics*.

Vermeersch, Christel (2002) "School Meals, Educational Achievement, and School Competition: Evidence from a Randomized Experiment," mimeo, Harvard University.