USING RANDOMIZED CONTROL DESIGNS IN EVALUATING SOCIAL SECTOR PROGRAMS IN DEVELOPING COUNTRIES

John Newman
Laura Rawlings
Paul Gertler

Seven case studies—from Bolivia, Colombia, Indonesia, Mexico, Nicaragua, Taiwan (China), and Turkey—demonstrate the feasibility of conducting rigorous impact evaluations in developing countries using randomized control designs. This experience, covering a wide variety of settings and social programs, offers lessons for task managers and policymakers interested in evaluating social sector investments.

The main conclusions are: first, policymakers interested in assessing the effectiveness of a project ought to consider a randomized control design because such evaluations not only are feasible but also yield the most robust results. Second, the acute resource constraints common in developing countries that often make program rationing unavoidable also present opportunities for adopting randomized control designs. Policymakers and program managers need to be alert to the opportunities for building randomized control designs into development programs right from the start of the project cycle because they, more than academic researchers or evaluation experts, are in the best position to ensure that opportunities for rigorous evaluations are exploited.

Despite the importance of knowing whether social programs work as intended, evaluations of social sector investments are still uncommon in developing countries. This neglect of evaluation handicaps the development community's ability to demonstrate what has been achieved and so to win political support, design more effective projects, and set priorities for resource allocation. Today, as more money than ever is flowing to the social

The World Bank Research Observer, vol. 9, no. 2 (July 1994), pp. 181–201

sectors governments and lending institutions are demanding value from that money. Evaluations can help make that happen by answering the critical question of how effective a particular social sector intervention is relative to other possible interventions.

This article and the companion article by Grossman lay out the issues of which policymakers and task managers need to be aware to build successful evaluation designs into their projects. Grossman's article describes the advantages, disadvantages, and limitations of the three main types of evaluation strategy—two quasi-experimental (reflexive and matched comparison) and one experimental (randomized control) (see table 1)—and reviews their use in social sector programs in the United States. Examples can be found in developing countries for each type of evaluation strategy discussed in Grossman's article.[1] Grossman expresses the view, generally shared by evaluation experts, that randomized control designs are the best evaluation strategy in technical terms but that in many situations it is not possible or appropriate to apply them.

This article examines the use of randomized control designs in developing countries and reaches two main conclusions. First, whenever a project is of sufficient interest to policymakers to warrant an impact evaluation, program designers ought to consider a randomized control design because this methodology yields the most robust results. Second, rigorous randomized control designs can often be built into a social sector program when acute resource constraints make rationing of services unavoidable. The second point is not new (Blum and Feachem 1983), but it may be salutary to remind policymakers and program managers that randomized control designs can often be built into

Table 1. *Evaluation Strategies*

Type	*Control group selection criteria*	*Pros*	*Cons*	*Frequency of use*
None	None	Very cheap	Nothing is learned	Very common
Reflexive	Program participants' behavior before the intervention	Cheap	Change in outcome may be due to other factors	Occasional
Matched comparison	Judgmental pairing	Better than random when target population is small	Results may not be generalizable	Occasional
Random	Random	Statistical; inferences can be drawn from result	Can be expensive	Rare

a social sector program at relatively low cost. Program managers, rather than academic researchers or evaluation experts, are in the best position to ensure that the opportunities for rigorous evaluation are exploited.

These opportunities present themselves whenever, for administrative or budgetary reasons, the number of eligible candidates exceeds the number of participants that the program is capable of serving. In developing countries there may not be enough resources to provide the program to all potential beneficiaries at once or even to all members of a high-priority group. Program managers frequently allocate scarce services by spreading resources evenly but thinly among eligible participants or by tightening the eligibility criteria until the number of people eligible matches the resources available. A common procedure is to rank each individual, community, or geographical area according to priorities set by the program, on the basis of such criteria as per capita income or the percentage of households with substandard housing. The cutoff point is then determined according to available funds. Tests are rarely done for the statistical significance of the differences in the indicators used in the ranking. Thus, it is entirely possible that individuals or communities that are observationally equivalent and equally eligible would be assigned different probabilities for receiving the program.

If all potential beneficiaries are equally eligible, a random draw can be used to select among them, and those who are not selected can serve as controls for those who are. This procedure need not be incompatible with targeting, since eligibility can be restricted to members of a high-priority group. The element of randomization ensures both equity in the allocation process and equivalence in the treatment and control groups.

Often, policymakers and program managers believe that conducting an impact evaluation of any type, especially one using a rigorous experimental design, would be too difficult or too costly in a developing country. In this article, we present a series of case studies that demonstrate that randomized control designs have been used successfully in developing countries and that no insurmountable barriers of knowledge, experience, or cost stand in the way of conducting such evaluations. We also point out some of the design and implementation issues that task managers may face when they try to implement rigorous evaluations in developing countries and note that such evaluations are not always warranted. In some cases, after weighing what could be learned from an evaluation against the costs of carrying it out, it may make sense to decide not to conduct an evaluation. Most published impact evaluations pay little attention to costs—both the costs of carrying out the intervention and those of conducting the evaluation. Whether the evaluations themselves share this shortcoming or whether the published reports merely fail to provide the information, the outcome is a dearth of published data on the cost of evaluation. In the conclusion, we discuss some issues related to costs and provide some practical suggestions on setting up randomized control evaluations in developing countries.

Randomized Control Designs Work in Developing Countries

This article presents seven success stories. The seven cases used randomized control design to evaluate the impact of social sector projects ranging from family planning to radio education and mass communication. Randomized control designs have been applied successfully in many diverse settings and programs in developing countries, although they have been used much less often than have other evaluation methodologies and much less often than they have been in industrial countries. Boruch, McSweeny, and Sonderstrom (1978) found that of 400 documented cases of randomized control designs in settings outside of laboratories, less than 5 percent were conducted in developing countries. A review by Cuca and Pierce (1977) found that only twelve of ninety-six family planning program evaluations used randomized control design.

Few impact evaluation studies of any type, but particularly those using randomized control designs, have been carried out in developing countries in recent years. This scarcity is reflected in the fact that few of our examples are drawn from the 1980s. This seeming reluctance to conduct evaluations sometimes appears to stem from a sense that such studies are too expensive and too complicated to justify their use. The real problem, however, may be that evaluations have been inappropriately applied. Policymakers and program managers may have been discouraged by efforts to evaluate program impacts when the programs themselves were suffering from severe implementation problems.[2] An impact evaluation is not the appropriate tool for monitoring whether a program is functioning as it was designed to function. That is the purpose of a monitoring system, which provides inexpensive and timely information on the program and beneficiaries and on whether the program is being implemented as intended. To determine whether a program, properly implemented, has the desired effect requires an evaluation strategy that, in addition, collects data from an appropriate comparison group. Monitoring programs can be simple and cheap—indeed, multilateral lending institutions are recommending that monitoring information be produced routinely in all projects that they finance. Evaluation is harder.

The advantage of a technically sound impact evaluation is that it can provide convincing evidence of program effectiveness for policymakers. That involves collecting information on a comparison group as well as the treatment group and applying a rigorous design to ensure that differences in outcomes result from the impact of the program rather than from measured or unmeasured differences between the treatment and control groups. The technical soundness of the design can be instrumental in convincing policymakers of the reliability of a study's findings. The first two case studies, from Nicaragua and Turkey, illustrate how the use of a randomized control design convinced policymakers of the effectiveness of new approaches to learning. The right design can also help policymakers choose among alternative program options, as illustrated by the Colombia and Taiwan (China) case studies.

The implementation of an evaluation in a developing country can be as important as its design. A program manager setting out to conduct an impact evaluation in a developing country is also something of a pioneer. Typically, there are no consulting firms to call on to carry out the evaluation, as there are in the United States. Political support for the evaluation may be weak or absent. Further, many of the same factors that can make implementing a project difficult—the rapid turnover of staff, political change, sporadic interruptions in cash flow—can make conducting an impact evaluation difficult.

At the same time, the budgetary and administrative constraints in developing countries that often make it impossible to reach all potential beneficiaries at once create opportunities for using randomization that are less often encountered in established market economies. The need to ration services and benefits means that a randomized control design can be built into a program's first implementation phases, as happened in the case of the education upgrading program in Bolivia. Evaluating the first part of a phased-in program presents an alternative to a pilot program, which may not accurately predict the effect of the full-fledged program because of differences in the way pilot and full programs are implemented, as illustrated by the experience with the "Sesame Street" program in Mexico. In addition, using a randomized control design in the first part of a program can build up valuable experience in conducting evaluations in developing countries, making it in many cases a more useful exercise than promoting expensive pilot programs.

It is noteworthy that in six of the seven case studies, the programs delivered services to a community rather than directly to individuals, a common practice in developing countries. The experimental conditions required for a randomized control group design are less likely to be contaminated in a society in which communities are relatively self-contained, as they tend to be in developing countries. (See Grossman in this volume for a discussion of contamination of the control group.)

Even when a program is delivered to communities, indicators at both the individual and community levels may be used to measure its impact. The individual comparisons provide more accurate measurements of the program's impact, but they are statistically more demanding. When programs directed at communities are evaluated using community-level variables, unbiased estimates of the impact of the availability of the program on measured community outcomes can be obtained without controlling statistically for the correlation between an individual's decision to participate in the program and the outcome. (See Grossman in this volume for a discussion of the problem of disentangling participation and treatment effects.) The use of communitywide averages combines the outcomes for individuals in the treatment community who choose not to participate in the program with those for individuals who do participate. Provided that a sufficiently large number of communities are included in the program and control groups, the measured differences in community-level indicators between the program and control areas would yield

estimates of the expected effect of extending the program to similar, unserved communities. The community-level differences would not, however, yield estimates of the potential impact of extending the program benefits to all individuals or to a target group of individuals.

A related problem is that most programs that require rationing are not assigned randomly to eligible communities, as they were in the Bolivia education upgrading project. Thus, differences in outcomes across communities may reflect a combination of the program's impact and an explicit or implicit allocation rule that may incorporate measured or unmeasured differences across communities. Failure to account for unmeasured differences that are related both to program allocation and to outcomes can yield biased estimates of a program's impact. In projects that require communities or individuals to apply for services, it is especially important that the evaluation be designed to analyze both the decision to apply for services and the impact of the project.[3] For example, the Indonesia National Family Planning Coordination Board allocates more family planning resources to communities in which contraceptive prevalence is low. One study (Lerman and others 1989) reported a negative correlation between family planning program inputs and contraceptive prevalence using least squares cross-section multivariate regressions. However, this result says more about the effect of past contraceptive choices on the way the government allocates program inputs than it does about the effect of those inputs on couples' contraceptive choices.

Rosenzweig and Wolpin (1986) have pointed out that most of the economic studies that have attempted to evaluate social sector interventions have ignored this problem and have implicitly assumed that program managers randomly allocate programs across communities. They demonstrate that information over time on the spatial distribution of programs and program characteristics can be used to yield unbiased estimates of the effects of *changes* in local programs on *changes* in local population characteristics. Working with changes eliminates the influence that unmeasured, fixed characteristics of the community could have on the outcome.[4] Using repeated observations of program interventions and household outcomes in ex post matched comparisons is a promising approach that is worth pursuing. However, substantial improvements would have to be made in national information systems to generate and then link adequate information on program interventions (typically collected from community surveys, provider surveys, and administrative records) with household outcomes (obtained from household surveys) before any useful results could be realized. Even good national information systems are not yet designed so that these links can be made easily. The World Bank's Living Standards Measurement Study is encouraging further efforts along these lines.

Not all evaluations in developing countries will focus on the impact of expanding services to other groups or individuals. Evaluations have also been used to test the feasibility of introducing changes in the price of services delivered, as is illustrated by the case from Indonesia. The Indonesian case also

underscores some of the political constraints that can be encountered when applying randomized control designs and the tradeoffs that must often be made between these political constraints and the reliability of the evaluation design.

Showing That Radio Education Works: The Radio Mathematics Project in Nicaragua

This project used a randomized control evaluation design to assess and demonstrate the effectiveness of a new approach to learning—radio education. The positive findings of the evaluation led to the expansion of the radio education program to classrooms throughout Nicaragua and to further use of randomized control in evaluating the effectiveness of radio education compared with that of new textbooks.

The Radio Mathematics Project was launched in 1974 by Stanford University through the Ministry of Public Education, with the support of the U.S. Agency for International Development (USAID). The aim was to develop and implement a prototype system of radio-delivered mathematics instruction for elementary school students. The project was implemented in four phases—research, pilot-level field tests, standardized tests, and the main field test.

The first two years, 1974 and 1975, were dedicated to establishing the project, developing lessons, and conducting pilot tests of the program in first grade classrooms in California and in Masaya, Nicaragua. In 1976 and 1977 schools in the provinces of Masaya, Carazo, and Granada were randomly selected to receive the revised mathematics program. In 1978 the Province of Rio San Juan was also included in the project.

School populations were categorized by grade and by rural and urban areas in each province so that the effect of the program on different groups within the population could be assessed. Within each province each qualifying school (any school with at least fifteen first graders) had an equal chance of being in the treatment group or in the control group. Each year, depending on the grade being evaluated, schools were chosen from the list of randomly assigned treatment and control groups using a three-step process. First, the number of classes to be chosen from each group was determined. Next, a list of eligible classes was drawn up for each cell in each category (for example, rural control schools in Masaya). Finally, the appropriate number of classes was selected from each list. From 1975 to 1978, this process generated a total of 145 control classes and 257 treatment classes for the evaluation.

The radio education program for the first through fourth grades consisted of an hour of mathematics instruction daily throughout the school year, divided into radio instruction and teacher-assisted exercises. During the period when the program was being fully implemented (1976–78), project personnel

administered tests to students in the control and treatment groups both before and after the program aired.

Quantitative evaluations showed statistically significant improvements on mathematics tests for students in the first through third grades who received the radio education program (Friend, Searle, and Suppes 1980). For the first grade the mean correct score on the tests was 65.5 percent for the treatment classes, but only 38.8 percent for the control classes. In the second grade the scores were 66.1 and 58.4 percent, and in the third grade, 51.7 and 43.2 percent. For all three grades these differences in scores were statistically significant at the 99 percent level of confidence, that is, it is 99 percent likely that the differences between the control and treatment groups could be attributed to the treatment rather than to chance. Scores for fourth graders were not statistically different for the treatment and control classes, but this grade was tested during a period of revolutionary turmoil, when many schools dismissed children before the daily broadcast of the fourth grade lesson—an extreme example of how failure to implement a project as planned precludes meaningful evaluation.

Qualitative evaluations based on classroom observation and weekly tests also constituted an important part of the overall evaluation. These activities allowed teaching methods to be assessed and refined rapidly and provided valuable feedback to teachers. The qualitative evaluations found students to be attentive and able to keep pace with the worksheets and to learn new skills. Teachers reported satisfaction with the program, which they said reduced their workload and introduced students to new concepts.

Explicit efforts were made to build political support for the evaluation. Two advisory committees, with representatives from the Ministry of Public Education and participating schools, were established to explain the objectives of the program and the evaluation. Briefing sessions were conducted to explain the use of randomized control design and to reassure teachers that the program, not their teaching, was being evaluated. Each eligible school had the same probability of being selected to receive the program, lessening the chance of any school or individual developing feelings of animosity toward a program that had "rejected" them, which could have influenced the results.

This success led to a second evaluation using a randomized control design, which confirmed the greater effectiveness of the radio education program in increasing children's learning ability compared with a program that provided additional textbooks (Jamison and others 1981). Since this trial run in Nicaragua, the number of interactive radio-based education programs in developing countries has grown steadily. During the 1980s radio mathematics programs were introduced in Bolivia, Costa Rica, the Dominican Republic, Ecuador, Guatemala, Honduras, Lesotho, Nepal, and Thailand. Interactive radio instruction programs in science, health, Spanish, English as a Second Language (ESL), and teacher training have also spread across the developing world (USAID 1990).

Testing for Lasting Effects: Early Childhood Education in Turkey

In 1982 a pilot program headed by the Psychology Department at Bogazici University in Istanbul, Turkey, was initiated to test whether educating lower-income mothers of three- and five-year-olds improves the children's learning abilities. Because the beneficial effects of early-childhood interventions provided directly to children had often been found to dissipate with time, the program managers hoped that, by educating mothers instead of children, the program would have a lasting effect on children's cognitive abilities. The hypothesis was that the mothers' training would constitute a permanent change in the children's environment. This program was evaluated twice: once at the time of the project, to assess immediate effects; and again nine years later, to find out whether the effects were lasting—an ambitious follow-up program.

A series of assessments, tests, and interviews were used to establish a baseline for the project. Three categories of mothers were then selected to receive training: those whose children were attending an educational preschool, those whose children were attending a custodial daycare center, and those who were caring for their children at home. Treatment and control groups were established through random selection. The treatment group began a two-year, two-part training program that consisted of a cognitive development program for children, implemented through a series of exercises completed by mother and child working together, and an enrichment program that educated mothers about their children's health and education needs.

An initial impact assessment was conducted at the end of the two-year training program. The children of mothers who had gone through the program scored significantly higher in measures of IQ, analytical training, and classification tasks than children in the control group. They also had higher grades, most notably in Turkish and mathematics.

Because the initial evaluation showed such positive results, a revised version of the program was extended to other areas of the country, with the support of nongovernment organizations and private industry. A television version of the enrichment training for mothers was also developed and broadcast in a series of eleven short programs.

A long-term impact evaluation was recently completed for 217 of the original 255 participants in the training program—a follow-up rate of 85 percent (Kagitcibasi, Sunar, and Bekman 1993). The evaluation included interviews with the children, now twelve to fifteen years old, and their parents. The results of this study confirmed the hypothesis that changing the environment in which children learn can lead to sustainable improvements in education. One of the most striking long-term impacts of the training is the much higher school retention rates for the children whose mothers participated in the program: 86 percent, compared with 67 percent for the children of mothers in the control group. Throughout the first five years of primary school, the academic performance and vocabulary test scores of children whose mothers had

received the training were consistently superior to those of children whose mothers had not. In addition, both the children and the mothers who had benefited from the training program had significantly different scores for answers on questions that demonstrated self-confidence, attitudes toward academics, and expectations about educational achievement.

Testing Alternative Service Delivery Modes: The Taichung Family Planning Program in Taiwan, China

In 1962 the Taiwan Provincial Health Department began what was at the time the largest intensive family planning program ever carried out in a city the size of Taichung, which had a population of 325,000. The decision to extend the program to the entire city was prompted by the results of a series of surveys in 1961–62 that revealed a strong demand for family planning services and a readiness to use a new form of birth control, the intrauterine device (IUD). Information services and supplies were offered for a wide variety of contraceptive methods.

Program officials chose to test the effectiveness of different combinations of services and information by randomly assigning treatments by *lin*, a neighborhood unit averaging twenty households. In all, some 36,000 married couples of childbearing age (couples in which the wife was between the ages of twenty and thirty-nine) were included in control and treatment groups. Four types of treatment were designed, ranging from more intensive and more costly to less intensive and less costly:

- *Treatment 1: Everything, husband and wife.* Personal visits to husbands and wives by trained health workers providing family planning information and services; mailings to newlyweds and couples with at least two children detailing family planning methods and benefits and identifying the location of clinics; and neighborhood family planning meetings offering information about family planning.
- *Treatment 2: Everything, wife only.* Same as treatment 1, but without the visits to the husband by the health workers.
- *Treatment 3: Mailings.* Only informational mailings, as detailed in treatment 1.
- *No treatment.*

In addition, the city was divided into three "density" sectors, which differed, insofar as possible, only in the proportion of *lins* receiving more intensive or less intensive treatments. The density variation was introduced to determine to what extent the beneficiary population could be depended on to spread the desired innovation and to establish how many households within a given area needed to be contacted to stimulate diffusion of the innovation. Differences among the three density sectors were minimized by constructing sectors that were as similar as possible on the basis of measurable characteristics such as

Table 2. *Cumulative Acceptance Rates per 100 Married Women Aged 20–39 for All Methods of Birth Control in Taichung*
(percentage)

	Density sector			
Treatment	*Heavy*	*Medium*	*Light*	*All sectors*
Treatment 1	20	12	14	17
Treatment 2	18	14	14	17
Treatment 3	8	7	8	8
Nothing	9	7	7	8
Total	14	9	8	11

Source: Authors' calculations, from Freedman and Takeshita (1969).

fertility, occupational composition, and education. In the sector designated to receive high-density treatment (928 *lins*), half the couples were randomly chosen to receive an "everything" treatment (treatment 1 or 2). In the sector designated to receive low-density treatment (730 *lins*), only 20 percent of the couples received an "everything" treatment. In the medium-density sector, 34 percent of couples received an "everything" treatment. Each *lin* within each sector had the same probability of receiving a treatment because the treatments were allocated randomly by *lin*. However, the probability of being selected into each treatment category varied according to the treatment density to which the sector was assigned (Freedman and Takeshita 1969).

During the experimental period of the program from February 1963 to March 1964, the contraceptive acceptance rate was significantly higher in the high-density sector than in the medium- or low-density sectors (table 2). The variation between medium- and low-density sectors was slight. The experiment suggested that the marginal effect of approaching the husbands (treatment 1) in addition to the wives (treatment 2) was negligible and that the mail campaign (treatment 3) was largely ineffective. In 1964, elements of the Taichung program that were considered the most promising—notably house visits by fieldworkers—were replicated throughout Taiwan, and greater emphasis was placed on the availability of IUDs as a method of family planning.

Targeting and Random Assignment: The Cognitive Abilities of Malnourished Children in Colombia

A pilot program in Cali, Colombia, in 1971–75 was designed to determine what levels of education, nutrition, and health services for preschool children and parents from low-income families would reduce malnutrition and whether these actions could produce improvements in children's intellectual functioning (McKay and others 1978). Medical practitioners had long asserted that inadequate nutrition impairs a child's cognitive development, perhaps permanently,

but these claims had never been systematically investigated. This case shows that, when program services are to be phased in, a randomized control design can be used even in a program that aims eventually to cover all eligible participants. Random assignment is used simply to determine which groups or individuals receive the program first. This case also shows that achieving an efficient randomized control design may require that the target group be identified first.

The program was run by the staff of the Human Ecology Research Station, with the support of the Colombian Ministry of Education, the Ford Foundation, the National Institute for Child Health and Human Development, and a number of private industries in Colombia. The first step was a multiphase screening survey to identify a target group of malnourished children from among households with four-year-old children. The survey identified general nutritional levels, gathered demographic data, and screened for malnutrition. The 333 malnourished children identified through this process were classified into twenty sectors by neighborhood. Each sector of thirteen to nineteen children was randomly assigned to one of four treatment groups that differed only in the duration of the treatments, which were staggered over time. Two other groups of children of the same age were formed to allow for qualitative comparisons with the treatment groups. One group consisted of children from high-income families living in Cali, and the other of children from low-income families who exhibited no signs of malnutrition but who lived in the same neighborhoods and participated in the screening process that had identified the children who qualified for the program.

The children in the treatment groups participated in six hours of health- and nutrition-related and educational activities a day, five days a week. The nutritional component provided 75 percent of recommended daily protein and calorie intake, along with mineral and vitamin supplements. Health care services included daily observations of all children and immediate pediatric attention as warranted. The educational component focused on developing cognitive processes and language, social, and psychomotor skills.

Because one of the objectives of the study was to assess how long such a program should last, time-sequencing of treatments formed a crucial part of the pilot program. A randomly selected subgroup from the larger pool of malnourished children was assigned to treatment 4, the longest treatment period of 4,170 hours. Over staggered eight-month periods, other randomly selected subgroups received treatments 3, 2, and 1; the last was the shortest, lasting only 990 hours. The children's development was traced over the forty-four months of the program by measuring each child's cognitive ability at equally spaced intervals five times during the study period. The tests measured such indicators of cognitive ability as use of language, spatial relations, quantitative concepts, logical thinking, and manual dexterity and motor control. One problem, however, is that different tests were administered at each measurement point, making it difficult to compare the test results.

Because children were assigned randomly to the four treatment groups, differences among the groups could be attributed to differences in the duration of the program. Children who received the longest treatment showed the greatest gains. For children eight years old, results on the Stanford-Binet intelligence tests—reported as mental age minus chronological age—were as follows for the different groups: treatment 1, –15 months; treatment 2, –11 months; treatment 3, –9 months; and treatment 4, –5 months. The treatment groups differed from one another in the expected direction—the longer the treatment, the greater the gain—but the differences between adjacent treatment groups were not statistically significant. (It should be noted, however, that the sample sizes were small.) Even with the maximum treatment, none of the groups ever reached the average level of ability shown by children from the nonrandomly selected high socioeconomic group, who had a mental age minus chronological age of +10 months as measured by the Stanford-Binet test.

No member of the target group was denied treatment, a factor that facilitated acceptance of the randomized control design, particularly in the sensitive case of a study of the effects of malnutrition on intellectual development.

Testing the Whole Program: The Impact of "Sesame Street" in Mexico

A new version of the children's television program "Sesame Street," in Spanish and adapted to Latin American culture, was introduced in Mexico in 1971. Policymakers in the communications and education fields were interested in exploring the effect of the program on children's cognitive skills. The evaluation was designed to assess the effectiveness of the entire program, rather than the relative effectiveness of different strategies, as in some of the other case studies. This case illustrates some of the problems that can occur in moving from a pilot program to broader implementation of the project.

A randomized control design was applied to a pilot test carried out in daycare centers serving low-income families in Mexico City in 1971. Two hundred and twenty-one children three to five years old from three daycare centers were divided by age and gender and then randomly assigned to treatment or control groups. Children in the treatment group watched "Plaza Sesamo" for fifty minutes a day five days a week for six months. Children in the control group watched cartoons. To make sure that children in the control group did not watch "Plaza Sesamo" at home in the evening (it was shown again at 6:00 P.M.), children in that group were kept at the daycare centers until 7:00 P.M.; children in the treatment group left earlier (Hoole 1978).

Nine cognitive development tests were administered to the randomly selected control and experimental groups before and after the pilot program began. Statistically significant differences were found for four of the nine cognitive tests administered after the program. The greatest differences were

in the tests of letters and words, general knowledge, and numbers—topics most closely related to the objectives of "Plaza Sesamo." (Differences after the program in adjusted test mean scores for four- and five-year-olds in the experimental group and those in the control group were 7.3 and 4.8 in general knowledge; 4.5 and 5.1 for letters and words; and 7.8 and 6.2 in numbers, all significant at the 99 percent confidence level.)

The encouraging results of this pilot test prompted a larger field test. Control and treatment groups were randomly selected from lower- and middle-class preschool children in daycare centers in urban and rural areas. The impact of "Plaza Sesamo" was not as clear in the broad field test, which used a slightly different methodology (the tests were revised, and a rural component was added). The field test was also ultimately less rigorous because of the larger number of dropouts and contamination that occurred because some children had watched an earlier version of "Plaza Sesamo" at home. However, the evaluators suggested that the difference between the pilot test and the field experiment resulted less from the difference in methodology than from important differences in the social environments in which the children watched the program (Diaz-Guerrero and others 1976). Essentially, they hypothesized that the presence of a greater number of adults in the laboratory-type setting of the pilot project created a slightly different environment that was more conducive to learning. Because the laboratory-type setting was not replicated when the program was expanded, the nature of the intervention changed.

Although the results of the field test were less conclusive, the results of the pilot test helped to generate broad interest in "Sesame Street," not only in Mexico but throughout Latin America.

Assigning Services by Lottery: Educational Investments in the El Chaco Region of Bolivia

This case study and the malnourished children project in Colombia both illustrate that the targeting of project interventions does not have to rule out the use of randomized control designs for evaluation. When resources are limited, it may be preferable to group individuals or communities on the basis of some rough classification criteria, treat them as observationally equivalent, and conduct a lottery to distribute limited resources, rather than spend funds on more costly information collection activities to target services more narrowly. This approach was followed in a pilot program recently introduced in the El Chaco region of Bolivia to upgrade physical facilities and teacher training in rural public schools (Coa 1992). The program is one of several activities financed by the Social Investment Fund (SIF)—an institution set up by the Government of Bolivia to finance education and health projects in low-income areas—and is also supported by the World Bank and Kreditanstalt für Wiederaufbau.

To direct interventions to the neediest cases, project managers assigned schools in the region to one of three priority groups on the basis of community characteristics and assessments of the current state of their infrastructure. Recognizing that funds spent making finer distinctions among schools could be better spent on program activities, project managers made no attempt to measure subtle differences among the schools or to rank them in order of priority.

All eight schools in the highest priority group were upgraded under the project. The next highest priority group contained 120 schools, but funds were available to upgrade only 54 of them. These schools were selected randomly. This group is of particular interest to policymakers because schools in this category are the hardest hit by current budget stringencies.

Because the allocation rule assigned services to all the schools in the top priority group, the effect of the intervention on that group will be measured using a reflexive comparison design (see the Grossman article for a discussion of this type of design). For the medium-priority group, conditions are right for using a randomized control group design. Baseline information for the evaluation was collected between May and June 1993 using household, community, and school facility questionnaires. A follow-up survey will be conducted one year later, after the project interventions have been completed.[5]

Combining Randomized Control and Matched Comparisons: The Indonesia Resource Mobilization Study

Sometimes evaluations combine randomized control group and matched comparison designs, as in the Indonesia Resource Mobilization Study. This study was designed in 1991 to ascertain the potential impact of user fees on health system revenues, health care utilization, patient's choice of medical care, provider services, and health outcomes and to assess the willingness of patients to pay for improvements in the health care system.

The Resource Mobilization Study is one component of the Third Health Project, a set of health care initiatives implemented by the government in the provinces of Kalimantan Timur and Nusa Tenggara Barat to increase the availability and improve the quality of medical services primarily through resource investments (such as new facilities, additional personnel, and more drugs and other medical supplies; see Indonesia 1992). The project has so far been funded by a World Bank loan, but unless the government finds other sources of financing once the loan is expended, the improvements in health care services will not be sustained. For that reason, the government wanted to take advantage of the opportunity presented by the health project to experiment with increases in user fees in two provinces before extending the scheme nationwide. The increases were likely to be less unpopular if they came at the same time as an overall improvement in the quality of services under the health project.

The interaction between the evaluation team and government policymakers led to several important and practical compromises in the design of the experiment. The government initially planned to increase fees uniformly across the two provinces, while expanding mechanisms to exempt the poor from having to pay the new fees. The evaluation team argued for delaying some of the fee increases so that experimental control and treatment groups could be studied. Random assignment of the fee increases at the individual level was clearly not practical because health care services are priced at the provider level. Applying the fee increase at the facility level would be difficult as well, because health care prices are set at the district level. Although local officials were eager to increase fees to generate additional revenue, they were reluctant to set different prices at different facilities within the same district for fear of political backlash. In the end, differences in fees were applied only at the district level, in six districts randomly selected from among the twelve in the two provinces. Fees were one and a half times higher than prevailing rates. Price variations were also introduced among levels of care (such as hospital and health center or inpatient and outpatient care).

The small size of the sample of districts subjected to fee increases created statistical problems, so a matched comparison was introduced to strengthen the evaluation design. Treatment and comparison villages were matched not directly on a village-by-village basis, but by comparing the distribution of socioeconomic characteristics of treatment and control villages as groups. First, 110 treatment villages were selected randomly from among the six randomly selected districts. Next, the same number of control villages was selected randomly from among the randomly selected control districts, and the distribution of their socioeconomic characteristics (income level, family size, access to medical care, and other data from national household surveys) was compared with that of the treatment villages. The control village that was the least similar to the treatment villages was dropped in favor of another randomly selected replacement village drawn from the control district, and the process was repeated until the comparability of the two groups could no longer be improved. This iterative process—made possible by the availability of national survey and census data on household- and village-level socioeconomic characteristics—substantially improved the fit of the match in one of the two provinces.[6]

Baseline information was collected in 1991 on the matched treatment and control villages using household, community, and health provider questionnaires. Follow-up surveys of the same households and providers were conducted in 1993, some eighteen to twenty months after the fee increases. Results of the analysis are expected in the summer of 1994. The collection of data both before and after the fee increases is intended to isolate the effect of the policy reforms from other factors that may have influenced people's use of medical services over time. The control-and-treatment-group design controls for other influences, such as changes in weather, morbidity pattern, and income, that

cannot be controlled for in a reflexive design, which tests the same group before and after the intervention.

Conclusion

These cases demonstrate the feasibility of conducting impact evaluations using randomized control designs in developing countries. They also demonstrate that there is no single blueprint for conducting evaluations. The evaluation designs explored in this article were tailored to the question of interest in the social sector project or treatment being evaluated. Such evaluations are most effective when they seek to answer a clear question of interest to policymakers and when the intervention itself can be precisely defined and measured: Can radio education improve learning? What is the most effective level of intensity in the provision of family planning services? How will people react to an increase in prices for medical services?

More effort needs to be devoted to collecting and reporting information on the costs of carrying out specific interventions. Having that information would allow the outcomes of different kinds of interventions to be expressed in terms of how much they cost to implement rather than in terms of outcome indicators that are not directly comparable. Because initial conditions and service delivery levels are often very poor in developing countries, an impact evaluation might easily find a sizable absolute improvement in the outcome indicators for given inputs. But it is important to remember that the relevant factor in deciding resource allocation is the opportunity cost of investing in one project rather than another; that is, the expected gains from investing in one project compared with the expected gains from investing in another.

More effort also needs to be devoted to collecting information on the costs of conducting evaluation studies. The critical question in deciding whether to conduct an evaluation is whether the expected value of the information obtained is greater than the cost of collecting it. Again, the relevant cost is the opportunity cost of using the funds. If the project to be evaluated is only one of a group of projects expected to have high returns with low risk, the opportunity cost of financing an impact evaluation instead of investing in another project might be high. If the level of uncertainty about what can be gained from the project is appreciable, however, spending the money on an evaluation study probably makes sense. The opportunity cost of investing in a project with low returns can be considerable when other investments could yield higher returns.

In addition to concern about the costs of conducting impact evaluations, policymakers and program managers need to be aware of some of the issues involved in setting up an impact evaluation study within a project. Some of the decisions made early on in the design of a project can make an impact evaluation easier or harder to conduct later: Who is eligible to participate in the

project? How are the project activities rationed among eligible beneficiaries if resources do not permit delivery of services to all who are eligible? How is the project being phased in? Policymakers and program managers should be alert to opportunities for introducing randomization into program implementation, thus building in possibilities for generating randomized control designs. Randomization can be used to allocate a limited number of spaces among equally eligible potential participants, as in the radio education project in Nicaragua. The education upgrading project in Bolivia shows that such opportunistic randomization need not be incompatible with targeting interventions to high-priority groups. Randomization may also be built into the plans for expanding a program: the last groups of participants to receive the program's benefits can serve as controls for the first groups. This approach is particularly appropriate in situations where it is ethically untenable to generate a control group that will be denied access to the program altogether. The Colombia case study of the malnourished children project is a good example of the use of this type of randomized control design.

In developing countries, the task of organizing an impact evaluation usually falls on program managers. It is rare to find either government agencies that have the capacity to conduct evaluations or local consulting firms that can be contracted to do the work. One way around some of these problems is for program managers to establish a small evaluation unit, preferably within the project unit. Household data collection can usually be subcontracted from a national statistical institute or a private company. Data on the internal operation of the project, including cost data and monitoring indicators, should be collected as part of the project's management information system. The evaluation unit should ensure that data on households, which will provide information on the outcomes, can be easily linked with the data on project inputs. Freeing personnel in the evaluation unit from direct data collection tasks allows them to concentrate on analyzing the data and bringing the results to the attention of program managers.

For some tasks, such as designing the evaluation and analyzing the data, the evaluation unit may need to call on consultants or technical assistance from lending institutions.[7] As the Indonesia case illustrates, there are often tradeoffs in the evaluation design that need to be analyzed by experts. The evaluation unit will also require support in addressing some of the conceptual issues involved in analyzing the data, particularly if the evaluation design relies on statistically controlling for differences between participants and nonparticipants in measuring impacts. The wide availability of powerful and cheap microcomputers and of user-friendly statistical software makes the task of processing the data much easier and cheaper than in the past.

By demonstrating a project's benefits, impact evaluations can also help to build political support for a project. Impact evaluations can also identify the best ways to carry out particular kinds of interventions and provide convincing evidence for changing or eliminating unsuccessful programs or components,

thereby improving the cost-effectiveness of project interventions. As the development community embarks on a major increase in social sector spending, it should reconsider the role that impact evaluations can play in ensuring the continual improvement of the quality of social sector investments. Only policymakers have the power to draw together all the parties involved in a planned intervention, allowing them to debate the merits of conducting an evaluation and of how best to proceed should they decide that evaluation is warranted. Policymakers and program managers need to be aware of the tradeoffs and feasibility of the various evaluation options before they can make an informed judgment.

Notes

John Newman is senior economist in the World Bank's Human Resources Division for Mexico and Latin America. Laura Rawlings is a consultant to the Poverty and Human Resources Division of the World Bank's Policy Research Department. Paul Gertler is senior economist at the RAND Corporation.

1. The most common form of evaluation in developing countries, as in industrial countries, is the matched comparison study. Examples of influential matched comparisons conducted in developing countries include those of television-based educational reform in El Salvador (Mayo, Hornick, and McAnany 1976), the Dacca family planning project in Pakistan, the Rajastan applied nutrition program in India (UNESCO 1984), and the Matlab family planning project in Bangladesh (Nag 1992; Balk and others 1988). A recent matched comparison is Revenga, Riboud, and Tan (1994) on employment programs in Mexico.

2. Berg (1987) and Binnendjik (1989) discuss some common concerns voiced about impact evaluation studies.

3. For further discussion of the problems involved in disentangling participation and impact, see the Grossman paper in this volume, Heckman (1992), and Manski and Garfinkel (1992).

4. Programs in Indonesia have been the subject of several evaluations that statistically control for the nonrandom placement of programs. Pitt, Rosenzweig, and Gibbons (1993) evaluated the impact of health and education programs on illness rates and school enrollment; Frankenberg (1993) evaluated the impact of health infrastructure on infant mortality; and Gertler and Molyneaux (1994) evaluated the impact of family planning programs on contraceptive prevalence and fertility.

5. The cost of collecting the data for the baseline and follow-up surveys in the El Chaco area is roughly US$300,000, about 0.4 percent of the total SIF budget of $74.5 million as of May 1993.

6. Both the sample size and the size of the fee increases were selected to obtain a statistical power of more than 80 percent. Power calculations used the national household survey data on health care utilization.

7. For practical information on conducting evaluations, see the "Program Evaluation Kit" put out by Sage Publications, Newbury Park, California, in 1987, which includes books on designing and implementing evaluations. Hoole (1978); Dennis and Boruch (1989); North (1988); and Freeman, Rossi, and Wright (1980) provide useful sources for exploring the developing-country context. For general information on evaluations, *Evaluation Review* may be consulted. For information on evaluation designs, the classic work by Campbell and Stanley (1963) is recommended. Fitz-Gibbon and Lyons Morris (1987) also provide practical information on designing evaluations. Rieken and Boruch (1984) provide further discussion of experimental designs in evaluating social programs.

References

The word "processed" describes informally reproduced works that may not be commonly available through library systems.

Balk, Deborah, Khodezatul Faiz, Ubaidur Rob, J. Chakraborty, and George Simmons. 1988. "An Analysis of Costs and Cost-Effectiveness of the Family Planning-Health Services Project in Matlab, Bangladesh." International Center for Diarrheal Research, Bangladesh. Processed.

Berg, Alan. 1987. *Malnutrition: What Can be Done? Lessons from World Bank Experience.* Baltimore, Md.: Johns Hopkins University Press.

Binnendijk, Annette L. 1989. "Donor Agency Experience with the Monitoring and Evaluation of Development Projects." *Evaluation Review* 13(3):206–22.

Blum, Deborah, and Richard Feachem. 1983. "Measuring the Impact of Water Supply and Sanitation Investments on Diarrhoeal Diseases: Problems of Methodology." *International Journal of Epidemiology* 12(3):357–65.

Boruch, Robert, John McSweeny, and John Soderstrom. 1978. "Randomized Field Experiments for Program Planning, Development, and Evaluation: An Illustrative Bibliography." *Evaluation Quarterly* 2(4):655–95.

Campbell, Donald, and Julian Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research.* Chicago, Ill.: Rand McNally.

Coa, Ramiro. 1992. "Diseño para la Evaluación de Impacto de las Intervenciones FIS." Fondo de Inversión Social, La Paz, Bolivia. Processed.

Cuca, Roberto, and Catherine Pierce. 1977. *Experiments in Family Planning: Lessons from the Developing World.* Baltimore, Md.: Johns Hopkins University Press.

Dennis, Michael, and Robert Boruch. 1989. "Randomized Experiments for Planning and Testing Projects in Developing Countries: Threshold Conditions." *Evaluation Review* 13(3):292–309.

Diaz-Guerrero, R., Isabel Reyes-Lagunes, Donald Witzke, and Wayne Holtzman, 1976. "Plaza Sesamo in Mexico: An Evaluation." *Journal of Communication* Spring:145–54.

Fitz-Gibbon, Carol Taylor, and Lynn Lyons Morris. 1987. *How to Design a Program Evaluation.* Newbury Park, Calif.: Sage Publications.

Frankenberg, Elizabeth. 1993. "The Effect of Access to Health Care on Infant Mortality in Indonesia: 1980–87." Dorothy S. Thomas Award Paper presented at 1993 Population Association of America (PAA) meetings, Cincinnati, Ohio. Processed.

Freedman, Ronald, and John Y. Takeshita. 1969. *Family Planning in Taiwan: An Experiment in Social Change.* Princeton, N.J.: Princeton University Press.

Freeman, Howard, Peter Rossi, and Sonia Wright, 1980. *Evaluating Social Projects in Developing Countries.* Paris: Organization for Economic Cooperation and Development (OECD).

Friend, Jamesine, Barbara Searle, and Patrick Suppes, eds. 1980. *Radio Mathematics in Nicaragua.* Stanford, Calif.: Stanford University Press.

Gertler, Paul, and John Molyneaux. 1994. "How Economic Development and Family Planning Programs Combined to Reduce Indonesian Fertility." *Demography* 21(1):33–64.

Heckman, James J. 1992. "Randomization and Social Policy Evaluation." In Charles Manski and Irwin Garfinkel, eds., *Evaluating Welfare and Training Programs.* Cambridge, Mass.: Harvard University Press.

Hoole, Francis W. 1978. *Evaluation Research and Development Activities.* Beverly Hills, Calif.: Sage Publications.

Indonesia, Ministry of Health. 1992. "Health Care Resource Needs and Mobilization in KalTim and NTB: Interim Results, Health Project III." WD-6281-1-MOH/RI. Jakarta.

Jamison, Dean, Barbara Searle, Klaus Galda, and Stephen P. Heyneman. 1981. "Improving Elementary Mathematics Education in Nicaragua: An Experimental Study of the Impact of Textbooks and Radio on Achievement." *Journal of Educational Psychology* 73(4):556–67.

Kagitcibasi, Cigdem, Diane Sunar, and Sevda Bekman. 1993. "Long-Term Effects of Early Intervention." Department of Education, Bogadzdi University, Istanbul, Turkey. Processed.

Lerman, Charles, John Molyneaux, Soetedjo Moeljodihardjo, and Sahala Pandjaitan. 1989. "The Correlation between Family Planning Program Inputs and Contraceptive Use in Indonesia." *Studies in Family Planning* 20(1):26–37.

Manski, Charles, and Irwin Garfinkel, eds. 1992. *Evaluating Welfare and Training Programs.* Cambridge, Mass.: Harvard University Press.

Mayo, J. K., R. C. Hornick, and E. G. McAnany. 1976. *Educational Reform with Television: The El Salvador Experience.* Palo Alto, Calif.: Stanford University Press.

McKay, H., A. McKay, L. Siniestra, H. Gomez, and P. Lloreda. 1978. "Improving Cognitive Ability in Chronically Deprived Children." *Science* 200(21):270–78

Nag, Moni. 1992. "Family Planning Success Stories in Bangladesh and India." Policy Research Working Paper 1041. World Bank, Population and Human Resources Department, Washington, D.C. Processed.

North, W. Haven. 1988. *Evaluation in Developing Countries: A Step in Dialogue.* Paris: OECD.

Pitt, Mark M., Mark R. Rozenzweig, and Donna M. Gibbons. 1993. "The Determinants and Consequences of the Placement of Government Programs in Indonesia: 1980–86." *The World Bank Economic Review* 7(3):319–48.

Revenga, Ana, Michelle Riboud, and Hong Tan. 1994. "The Impact of Mexico's Retraining Program on Employment and Wages." *The World Bank Economic Review* 8(2):247–77.

Rieken, Henry, and Robert Boruch. 1984. *Social Experimentation: A Method for Planning and Evaluating Social Intervention.* New York: Academy Press.

Rosenzweig, M., and K. Wolpin. 1986. "Evaluating the Effects of Optimally Distributed Public Programs: Child Health and Family Planning Interventions." *American Economic Review* 76(3):470–82.

Rossi, Peter, Howard Freeman, and Sonia Wright. 1979. *Evaluation: A Systematic Approach.* Newbury Park, Calif.: Sage Publications.

UNESCO (United Nations Educational, Scientific and Cultural Organization. 1984. *Project Evaluation: Problems of Methodology.* Paris.

USAID (U.S. Agency for International Development). 1990. *Interactive Radio Instruction: Confronting Crisis in Basic Education.* AID Science and Technology in Development Series. Washington, D.C.