
Annex 1

Case Studies

Annex 1.1: Evaluating the Gains to the Poor from Workfare: Argentina's TRABAJAR Program

I. Introduction

Project Description. Argentina's TRABAJAR program aims to reduce poverty by simultaneously generating employment opportunities for the poor and improving social infrastructure in poor communities. TRABAJAR I, a pilot program, was introduced in 1996 in response to a prevailing economic crisis and unemployment rates of over 17 percent. TRABAJAR II was launched in 1997 as an expanded and reformed version of the pilot program, and TRABAJAR III began approving projects in 1998. The program offers relatively low wages in order to attract ("self-select") only poor, unemployed workers as participants. The infrastructure projects that participants are hired to work on are proposed by local government and nongovernmental organizations (NGOs), which must cover the non-wage costs of the project. Projects are approved at the regional level according to central government guidelines.

The program has undergone changes in design and operating procedures informed by the evaluation process. TRABAJAR II included a number of reforms designed to improve project targeting. The central government's budget allocation system is now more heavily influenced by provincial poverty and unemployment indicators, and a higher weight is given to project proposals from poor areas under the project approval guidelines. At the local level, efforts have been made in both TRABAJAR II and III to strengthen the capability of provincial offices for helping poor areas mount projects and to raise standards of infrastructure quality.

Impact Evaluation. The evaluation effort began during project preparation for TRABAJAR II and is ongoing. The aim of the evaluation is to determine whether or not the program is achieving its policy goals and to indicate areas in which the program requires reform in order to maximize its effectiveness. The evaluation consists of a number of separate studies that assess (a) the net income gains that accrue to program participants, (b) the allocation of program resources across regions (targeting), (c) the quality of the infrastructure projects financed, and (d) the role of the community and NGOs in project outcome.

Two of the evaluation components stand out technically in demonstrating best practice empirical techniques. First, the study of net income gains illustrates best-practice techniques in matched comparison as well as resourceful use of existing national household survey data in conducting the matching exercise. Second, the study of targeting outcomes presents a new technique for evaluating targeting when the incidence of public spending at the local level is unobserved. The overall evaluation design also presents a best-practice mix of components and research techniques—from quantitative analysis to engineering site visits to social assessment—which provide an integrated stream of results in a timely manner.

II. Evaluation Design

The TRABAJAR evaluation includes an array of components designed to assess how well the program is achieving its policy objectives. The first component draws on household survey data to assess the income gains to TRABAJAR participants. This study improves upon conventional assessments of workfare programs, which typically measure participants' income gains as simply their *gross* wages earned, by estimating net income gains. Using recent advances in matched-comparison techniques, the study accounts for forgone income (income given up by participants in joining the TRABAJAR program), which results in a more accurate, lower estimate of the net income gains to participants. The second component monitors the program's funding allocation (targeting), tracking changes over time as a result of reform. Through judicious use of commonly available data (program funding allocations across provinces and a national census), the design of this component presents a new methodology for assessing poverty targeting when there is no actual data on program incidence. This analysis began with the first supervisory mission (November 1997) and has been updated twice yearly since then.

Additional evaluation components include a cost-benefit analysis conducted for a subsample of infrastructure projects, along with social assessments designed to provide feedback on project implementation. Each of these activities has been conducted twice, for both TRABAJAR II and TRABAJAR III. Three future evaluation activities are planned. The matched-comparison research technique will be applied again to assess the impact of TRABAJAR program participation on labor market activity. Infrastructure project quality will be reassessed, this time for projects that have been completed for at least one year to evaluate durability, maintenance, and utilization rates. Finally, a qualitative research component will investigate program operations and procedures by interviewing staff members in agencies that sponsor projects as well as program beneficiaries.

III. Data Collection and Analysis Techniques

The assessment of net income gains to program participants draws on two data sources, a national living standards survey (Encuesta de Desarrollo Social—EDS) and a survey of TRABAJAR participants conducted specifically for the purposes of evaluation. (The EDS survey was financed under another World Bank project. It was designed to improve the quality of information on household welfare in Argentina, particularly in the area of access to social services and government social programs.) These surveys were conducted in August (EDS) and September (TRABAJAR participant survey) of 1997 by the national statistical office, using the same questionnaire and same interview teams. The sample for the EDS survey covers 85 percent of the national population, omitting some rural areas and very small communities. The sample for the TRABAJAR participant survey is drawn from a random sample of TRABAJAR II projects located within the EDS sample frame and generates data for 2,802 current program participants (total TRABAJAR II participants between May 1997 and January 1998 numbered 65,321). The reliability of the matching technique is enhanced by the ability to apply the same questionnaire to both participants and the control group at the same time and to ensure that both groups are from the same economic environment.

To generate the matching control group from the EDS survey, the study uses a technique called propensity scoring. (The fact that the EDS questionnaire is very comprehensive, collecting detailed data on household characteristics that help predict program participation, facilitates the use of the propensity scoring technique.) An ideal match would be two individuals, one in the participant sample and one in the control group, for whom all of these variables (x) predicting program participation are identical. The standard problem in matching is that this is impractical given the large number of variables contained in x . However, matches can be calculated on each individual's propensity score, which is simply the probability of participating conditional on (x). (The propensity score is calculated for each observation in the participant and control group sample by using standard logit models.) Data on incomes in the matching control group of nonparticipants allows the income forgone by actual TRABAJAR II participants to be estimated. Net income arising from program participation is then calculated as total program wages minus forgone income.

The targeting analysis is remarkable in that no special data collection was necessary. Empirical work draws on data from the ministry's project office on funding allocations by geographic department for TRABAJAR I (March 1996 to April 1997) and the first six months of TRABAJAR II (May to October 1997). It also draws on a poverty index for each department

(of which there are 510), calculated from the 1991 census as the proportion of households with “unmet basic needs.” This is a composite index representing residential crowding, sanitation facilities, housing quality, educational attainment of adults, school enrollment of children, employment, and dependency (ratio of working to nonworking family members). The index is somewhat dated, although this has the advantage of the departmental poverty measure being exogenous to (not influenced by) TRABAJAR interventions. To analyze targeting incidence, data on public spending by geographic area—in this case, department—are regressed on corresponding geographic poverty rates. The resulting coefficient consistently estimates a “targeting differential” given by the difference between the program’s average allocations to the poor and non-poor. This national targeting differential can then be decomposed to assess the contribution of the central government’s targeting mechanism (funding allocations across departments) versus targeting at the provincial level of local government.

The cost-benefit analysis was conducted by a civil engineer, who conducted a two-stage study of TRABAJAR infrastructure projects. In the first stage she visited a sample of 50 completed TRABAJAR I projects and rated them based on indicators in six categories: technical, institutional, environmental, socioeconomic, supervision, and operations and maintenance. Projects were then given an overall quality rating according to a point system, and cost-benefit analyses were performed where appropriate (not for schools or health centers). A similar follow-up study of 120 TRABAJAR II projects was conducted a year later, tracking the impact of reforms on infrastructure quality.

The social assessments were conducted during project preparation for both TRABAJAR I and TRABAJAR II. They provide feedback on project implementation issues such as the role of NGOs, the availability of technical assistance in project preparation and construction, and the selection of beneficiaries. Both social assessments were carried out by sociologists by means of focus groups and interviews.

IV. Results

Taking account of forgone income is important to gaining an accurate portrayal of workfare program benefits. Descriptive statistics for TRABAJAR II participants suggest that without access to the program (per capita family income minus program wages) about 85 percent of program participants would fall in the bottom 20 percent of the national income distribution—and would therefore be classified as poor in Argentina. However, matching-method estimates of forgone income are sizable, so that average net income gained through program participation is about

half of the TRABAJAR wage. Program participants could not afford to be unemployed in the absence of the program; hence some income is forgone through program participation. It is this forgone income that is estimated by observing the incomes of nonparticipants “matched” to those of program participants. However, even allowing for forgone income, the distribution of gains is decidedly pro-poor, with 80 percent of program participants falling in the bottom 20 percent of the income distribution. Female participation in the program is low (15 percent), but net income gains are virtually identical for male and female TRABAJAR participants; younger participants do obtain significantly lower income gains.

Targeting performance improved markedly as a result of TRABAJAR II reforms. There was a sevenfold increase in the implicit allocation of resources to poor households between TRABAJAR I and TRABAJAR II. One-third of this improvement results from better targeting at the central level, and two-thirds results from improved targeting at the provincial level. There are, however, significant differences in targeting outcomes between provinces. A department with 40 percent of people classified as poor can expect to receive anywhere from zero to five times the mean departmental allocation, depending upon the province to which it belongs. Furthermore, those targeting performance tended to be worse in the poorest provinces.

Infrastructure project quality was found to be adequate, but TRABAJAR II reforms, disappointingly, did not result in significant improvements. Part of the reason was the sharp expansion of the program, which made it difficult for the program to meet some of the operational standards that had been specified *ex ante*. However, projects were better at meeting the priority needs of the community. The social assessment uncovered a need for better technical assistance to NGOs and rural municipalities as well as greater publicity and transparency of information about the TRABAJAR program.

V. Policy Application

The evaluation results provide clear evidence that the TRABAJAR program participants do come largely from among the poor. Self-selection of participants by offering low wages is a strategy that works in Argentina, and participants do experience income gains as a result of participation (although these net gains are lower than the gross wage, owing to income forgone). The program does not seem to discriminate against female participation. TRABAJAR II reforms have successfully enhanced geographic targeting outcomes—the program is now more successful at directing funds to poor areas; however, performance varies and is persistently weak in a few provinces that merit further policy attention. Finally, dis-

appointing results on infrastructure project quality have generated tremendous efforts by the project team to improve performance in this area by enhancing operating procedures—insisting on more site visits for evaluation and supervision, penalizing agencies with poor performance in project completion, and strengthening the evaluation manual.

VI. Evaluation Costs and Administration

Costs. The cost for carrying out the TRABAJAR survey (for the study of net income gains) and data processing was approximately \$350,000. The two evaluations of subproject quality (cost-benefit analysis) cost roughly \$10,000 each, as did the social assessments, bringing total expenditures on the evaluation to an estimated \$390,000.

Administration. The evaluation was designed by World Bank staff member Martin Ravallion and implemented jointly with the World Bank and the Argentinean project team. Throughout its different stages, the evaluation effort also required coordination with several local government agencies, including the statistical agency, the Ministry of Labor (including field offices), and the policy analysis division of the Ministry for Social Development.

VII. Lessons Learned

Importance of Accounting for Forgone Income in Assessing the Gains to Workfare. Forgone income represents a sizable proportion (about half) of the gross wage earned by workfare program participants in Argentina. The result suggests that conventional assessment methods (using only the gross wage) substantially overestimate income gains and hence also overestimate how poor participants would be in absence of the program.

Propensity-Score Matching Method. When the matched-comparison evaluation technique is used, propensity scores allow reliable matches to be drawn between a participant and nonparticipant (control group) sample.

Judicious Use of Existing National Data Sources. Often, existing data sources such as the national census or household survey can provide valuable input to evaluation efforts. Drawing on existing sources reduces the need for costly data collection for the sole purpose of evaluation. Innovative evaluation techniques can compensate for missing data, as the assessment of TRABAJAR's geographic targeting outcomes aptly illustrates.

Broad Range of Evaluation Components. The TRABAJAR evaluation design illustrates an effective mix of evaluation tools and techniques. Survey data analysis, site visits, and social assessments are all used to generate a wide range of results that provide valuable input to the project's effectiveness and pinpoint areas for reform.

Timeliness of Results. Many of the evaluation components were designed explicitly with the project cycle in mind, timed to generate results during project preparation stages so that results could effectively be used to inform policy. Several components now generate data regularly in a continuous process of project monitoring.

VIII. Sources

Jalan, Jyotsna, and Martin Ravallion. 1999. "Income Gains from Workfare and Their Distribution." World Bank, Washington, D.C. Processed.

Ravallion, Martin. 1999. Monitoring Targeting Performance When Decentralized Allocations to the Poor Are Unobserved." World Bank, Washington, D.C. Processed.

Annex 1.2: Does Microfinance Really Help the Poor? New Evidence from Flagship Programs in Bangladesh

I. Introduction

Project Description. The microfinance programs of the Grameen Bank, the Bangladesh Rural Advancement Committee, and the Bangladesh Rural Development Board are flagship programs for those instituted in many other countries. These programs provide small loans to poor households who own less than one-half acre of land. Loans are accompanied by innovative contracts and loan schedules. The programs have served over 4 million poor clients in Bangladesh and have apparently been quite successful. For example, the top quartile of borrowers from the Grameen Bank consume 15 percent more and have almost twice as high a proportion of sons in school and a substantially increased proportion of daughters in school compared with the bottom quartile.

Highlights of Evaluation. The evaluation investigates the impact of the programs on 1,800 households in Bangladesh and compares them with a control group of households in areas without any microcredit financing. The major contribution of the study is to demonstrate that simple estimates of the impact of programs can be substantially overstated: correction for selection bias nullifies apparently impressive gains. The evaluation shows that much of the perceived gains is driven by differences in who gets the loans: they tend to be wealthier and work more than control groups. Once appropriate techniques are used, there is no impact of borrowing on consumption, and children in program areas actually do worse than children in control areas. The key determining factor is the fact that program lending has not followed eligibility guidelines—in fact, many of the borrowers had landholdings in excess of the half-acre maximum.

The evaluation both uses an interesting survey technique and makes imaginative use of econometric techniques. Another interesting angle is that the evaluation also looks at the effect of the impact on the variance as well as the mean outcome and finds that the main gain from the programs is risk reduction rather than increasing mean outcomes.

II. Research Questions and Evaluation Design

The researchers are interested in identifying the impact of microfinance programs on log consumption per capita, variance of log consumption, log labor per adult in previous month, variance of per adult log labor,

adult male labor hours in past month, adult female labor hours in past month, percentage of male school enrollment (ages 5 to 17), and percentage of female school enrollment (ages 5 to 17).

The evaluation is survey-based and covers 87 villages surveyed three times during 1991 and 1992. Villages were chosen randomly from a census and administrative lists, from 5 subdistricts that served as controls and 24 subdistricts where the programs were implemented. Twenty households were surveyed per village.

This enabled the researchers to split the households into five different types, depending on the eligibility criterion of holding one-half acre of land. It is worth reproducing the schematic, which illustrates how to create dummy variables that characterize the typology and how to think about selection bias.

Village 1: With program			Village 2: Control
A Not eligible [b=1;e=0;c=0]		Households with more than 1/2 acre	B would not be eligible [b=0;e=0;c=0]
C eligible but does not participate [b=1;e=1;c=0]	D Participants [b=1;e=1;c=1]		E Would be eligible [b=0;e=1;c=0]
		Households with 1/2 acre and below	

Comparing outcomes for group D with those for group C is fraught with selection problems: evidence suggests that group C households do not participate because they are afraid of not being able to pay back. If landholding is exogenous, groups C and D can be compared with group E, however, because outcome difference depends on program placement rather than self-selection. This is not true, of course, if there are differences across villages. If there are differences (due, possibly, to nonrandom placement), then it is better to take a difference-in-difference approach. Thus, an evaluator can calculate mean outcomes for C and D, mean outcomes for A, and then calculate the difference. Similarly, the difference between mean outcomes for E and mean outcomes for B can be calculated, and then the within-village differences can be compared.

III. Data

The researchers collected data on 1,798 households; 1,538 of these were eligible to participate and 905 actually participated. The surveys were col-

lected in 1991 and 1992 after the harvests of the three main rice seasons. The key variables of interest were consumption per capita in the previous week, the amount of credit received, amount of land held, labor supply in the past month, and demographic characteristics. A secondary data source on land transactions is also used to check on market activity in land.

IV. Econometric Techniques

There are three interesting components to the techniques used. The first is the use of administrative data to check the key assumptions necessary to use a regression discontinuity design strategy: the exogeneity of landholding. The second is a very nice use of nonparametric graphing techniques to describe both the probability of being found eligible and the probability of getting a loan as a function of landholdings. This is combined with a very good discussion of when it is appropriate to use a regression discontinuity design—since the graphical analysis suggests that there is no clear breaking point at 0.5 acre. Finally, the study primarily uses difference and difference-in-differences techniques.

V. Who Carried It Out

The data were collected by the Bangladesh Institute for Development Studies on behalf of the World Bank. The analysis was performed by researcher Jonathan Morduch.

VI. Results

The results suggest that almost all the apparent gains from the program are due to selection bias resulting from loan mistargeting. In particular, the authors find that 20 to 30 percent of the borrowers own more land than the half-acre maximum requirement for the program, which suggests that program officers are likely to bend the rules in unobservable ways. When the comparisons are restricted to only those borrowers who meet the land restriction, the authors find that average consumption in the villages with access to microfinancing is less than the controls with both the difference and difference-in-differences methods. This suggests that there was substantial mistargeting of program funds, and as a result regression discontinuity approaches cannot be used to analyze program effects.

The evaluation is also useful in the comparison of results from different econometric techniques: results differ markedly when fixed effects and difference-in-differences or simple difference approaches are used.

The evaluation makes a convincing case that the former is less appropriate when unobservable target group differences are used in making the location decision. However, there are conflicting results in the two approaches about whether the programs reduced variation in consumption and income, highlighting the need for longitudinal data. The impact on education is actually reverse after correction for selection bias.

It is also worth noting that although this analysis shows little impact of the treatment relative to the control group, the control group may not, in fact, have lacked access to financing because this may be supplied by NGOs. The expenditure of millions of dollars to subsidize microfinance programs is, however, called into question.

VII. Lessons Learned

There are several very important lessons from this study. The first is the importance of checking whether the program functions as prescribed. The second is the consideration of the appropriateness of regression discontinuity design versus difference in differences or simple difference techniques. The third is considering the impact of an intervention on the second as well as on the first moment of the distribution, since the reduction in risk may, in itself, be a useful outcome. There is a more fundamental lesson that is not directly addressed but is clearly learned from the study. That lesson is one of political economy: if there is a strong incentive to bend the rules, those rules will be bent.

VIII. Sources

Morduch, Jonathan. 1998. "Does Microfinance Really Help the Poor? New Evidence from Flagship Programs in Bangladesh." Processed, June 17.

Also see:

Khandker, Shahidur R. 1998. *Fighting Poverty with Microcredit: Experience in Bangladesh*. New York: Oxford University Press for the World Bank.

Annex 1.3: Bangladesh Food for Education: Evaluating a Targeted Social Program When Placement Is Decentralized

I. Introduction

Project Description. The Food for Education (FFE) program in Bangladesh was designed to increase primary school attendance by providing rice or wheat to selected households as an incentive to parents. This began as a pilot program but has grown in size and importance: its share of the Primary and Mass Education Division's budget grew from 11 percent in 1993–94 to 26 percent in 1995–96 and reached 2.2 million children, or 13 percent of total enrollment. The design is quite interesting: the program was hierarchically targeted in that FFE was given to all schools in selected economically backward geographic units with low schooling levels. Then households were chosen to receive the food by community groups within the geographic units, based on set, albeit somewhat discretionary, criteria (landless households, female-headed households, and low-income households). Children in these households must attend at least 85 percent of the classes each month.

Highlights of Evaluation. This evaluation is extremely useful because it illustrates what can be done when the intervention design is not at all conducive to standard evaluation techniques and when the evaluation has to be done using existing data sources. In fact, the approach in the FFE was almost the polar opposite to a completely random assignment: not only were the geographic areas chosen because they had certain characteristics but the individuals within them were chosen because they needed help. Thus, since the program was targeted at the poorest of the poor, simple analysis will understate the impact.

This intervention design creates a major problem with creating a counterfactual because clearly selection into the program is determined by the household's need for the program. The evaluation provides an innovative—and readily generalizable—approach to addressing the resulting bias by relying on the decentralization of the decisionmaking process. In brief, because the central government allocates expenditures across geographic areas, but local agents make the within-area allocation, the evaluation uses instrumental variable techniques based on geography to reduce the bias inherent in the endogenous selection procedure. The application of the method results in much higher estimated impacts of FFE than ordinary least squares approaches.

II. Research Questions and Evaluation Design

The research question is to quantify the impact of the FFE on school attendance, measured as the attendance rate for each household. There is little in the way of prospective evaluation design: the evaluation is performed with already existing data— in particular, using both a nationally representative household expenditure survey and a detailed community survey. The retrospective evaluation was in fact designed to obviate the need for a baseline survey; the evaluation simply needed surveys that included household characteristics and specific geographic characteristics of the household area. The subsequent sections provide more detail on how these can be structured so that they reliably infer the impact of the intervention.

III. Data

The data are from the 1995–96 Household Expenditure Survey (HES), a nationally representative survey conducted by the Bangladesh Bureau of Statistics that both includes questions on FFE participation and has a local level survey component. The authors use responses on demographic household characteristics, land ownership, school, and program variables from 3,625 rural households to identify the impact on school attendance. School attendance for each child is actually directly measured in the HES: both the days that are missed and the days that the school is closed are counted. The dependent variable was constructed to be the household average number of days school was attended as a proportion of the feasible number of days. Both parts of this survey are critical. On the one hand, information on the household helps to capture the impact of demographic characteristics on school attendance. On the other hand, information on the characteristics of geographic location helps to model the decisionmaking strategy of the centralized government and reduce the selection bias noted above.

IV. Econometric Techniques

The evaluation addresses two very important problems faced by field researchers. One is that program placement is decentralized, and hence the allocation decision is conditioned on variables that are unobservable to the econometrician but observable to the people making the decision. This means that the evaluation requires a measure that determines program placement at the individual level but is not correlated with the error term (and hence program outcomes). The second is that there is only a single cross-section survey to work with, with no baseline survey

of the participants, so it is difficult to estimate the pure impact of the intervention.

The evaluation is extremely innovative in that it uses the two-step allocation process itself as an instrument. The key feature that is necessary in order to do this is that the cross-sectional data include both household characteristics and geographic characteristics. In this particular case, the model is as follows:

$$W_i = \alpha IP_i + \beta' X_i + \eta' Z_i + \mu_i \quad (1)$$

Here W is the individual's welfare outcome, X and Z include household and geographic characteristics, and IP , which is the individual's placement in the program, is correlated with the error term. Clearly, and of fundamental importance in the evaluation literature, least squares estimates of \forall will be biased.

The evaluation uses the geographic differences in placement as instruments for individual placement, because this is not correlated with the error term, as well as household characteristics. This then characterizes this relationship as

$$IP_i = \gamma GP_i + \pi' X_i + v_i \quad (2)$$

It is important to note here that it is critical that Z contains all the information that is used in making the geographic placement decision. In this case, the two sets of geographic variables are used. One set of geographic variables is fairly standard and actually directly affects attendance decisions in their own right: distance to school, type of school, and school quality variables. The second set has to do with the placement decision itself and, although long, is worth noting for illustrative purposes. The variables include land distribution; irrigation intensity; road quality; electrification; distance and time to local administration headquarters and to the capital; distance to health care and financial facilities; incidence of natural disasters; attitudes to women's employment, education, and family planning; average schooling levels of the head and spouse; majority religion of the village; and the population size of the village. These are calculated at the village level and appear to predict selection fairly well: a probit regression on a total of 166 villages resulted in a relatively good fit (a pseudo- R^2 of 0.55). This suggests that these variables do in fact capture overall placement.

This set of equations can then be modeled by using three-stage least squares and compared with the results from ordinary least squares regression.

V. Who Carried It Out

The evaluation was carried out by Martin Ravallion and Quentin Wodon of the World Bank as part of a long-term collaborative effort between the Bangladesh Bureau of Statistics and the Poverty Reduction and Economic Management Unit of the World Bank's South Asia Region.

VI. Results

There are clear differences in the two approaches: the estimated impact of FFE using the three-stage least squares approach was 66 percent higher than the ordinary least squares estimates without geographic controls and 49 percent higher than with the controls. In other words, simple estimates that only control for variation across households (ordinary least squares without geographic controls) will substantially *understate* the effectiveness of the program. Even including geographic controls to apparently control for geographic placement does not erase the attendant bias. In substantive terms, the average amount of grain in the program appeared to increase attendance by 24 percent when the method outlined above was used.

It is worth noting that the key factor to make this a valid approach is that enough variables are available to model the targeting decision and that these variables are close to those used by administrators. If there are still omitted variables, the results continue to be biased.

VII. Lessons Learned

Many evaluations do not have the luxury of designing a data collection strategy from the ground up, either because the evaluation was not an integral part of the project from the beginning, or simply for cost reasons. This is an important evaluation to study for two reasons. First, it documents the degree of bias that can occur if the wrong econometric approach is used. Second, it describes an econometrically valid way of estimating the impact of the intervention without the cost and time lag involved in a prospective evaluation.

VIII. Source

Ravallion and Wodon. 1998. *Evaluating a Targeted Social Program When Placement Is Decentralized*. Policy Research Working Paper 1945, World Bank, Washington, D.C.

Annex 1.4: Evaluating Bolivia's Social Investment Fund

I. Introduction

Project Description. The Bolivian Social Investment Fund (SIF) was established in 1991 as a financial institution promoting sustainable investment in the social sectors, notably health, education, and sanitation. The policy goal is to direct investments to areas that have been historically neglected by public service networks, notably poor communities. SIF funds are therefore allocated according to a municipal poverty index, but within municipalities the program is demand-driven, responding to community requests for projects at the local level. SIF operations were further decentralized in 1994, enhancing the role of sector ministries and municipal governments in project design and approval. The Bolivian SIF was the first institution of its kind in the world and has served as a prototype for similar funds that have since been introduced in Latin America, Africa, and Asia.

Impact Evaluation. Despite the widespread implementation of social funds in the 1990s, there have been few rigorous attempts to assess their impact on poverty reduction. The Bolivian SIF evaluation, carried out jointly by the World Bank and SIF, began in 1991 and is ongoing. The study features baseline (1993) and follow-up (1997) survey data that combine to allow a before-and-after impact assessment. It includes separate evaluations of education, health, and water projects and is unique in that it applies a range of evaluation techniques and examines the benefits and drawbacks of these alternative methodologies. The initial evaluation results are complete and are currently being presented to donors and government agencies for feedback. Final results and methodological issues will be explored in greater depth in conjunction with the Social Investment Funds 2000 report, along with an analysis of cost-effectiveness.

II. Evaluation Design

The Bolivian SIF evaluation process began in 1991, and is ongoing. The design includes separate evaluations of education, health, and water projects that assess the effectiveness of the program's targeting to the poor as well as the impact of its social service investments on desired community outcomes such as improved school enrollment rates, health conditions, and water availability. It illustrates best-practice techniques in evaluation

using baseline data in impact analysis. The evaluation is also innovative in that it applies two alternative evaluation methodologies—randomization and matched comparison—to the analysis of education projects and contrasts the results obtained according to each method. This is an important contribution because randomization (random selection of program beneficiaries within an eligible group) is widely viewed as the more statistically robust method, and yet matched comparison (using a nonrandom process to select a control group that most closely “matches” the characteristics of program beneficiaries) is more widely used in practice.

III. Data Collection and Analysis Techniques

Data collection efforts for the Bolivian SIF evaluation are extensive and include a pre-SIF II investment (“baseline”) survey conducted in 1993 and a follow-up survey in 1997. The surveys were applied to both the institutions that received SIF funding and the households and communities that benefit from the investments. Similar data were also collected from comparison (control group) institutions and households. The household survey gathers data on a range of characteristics, including consumption, access to basic services, and each household member’s health and education status. There are separate samples for health projects (4,155 households, 190 health centers), education projects (1,894 households, 156 schools), water projects (1,071 households, 18 water projects) and latrine projects (231 households, 15 projects).

The household survey consists of three subsamples: (a) a random sample of all households in rural Bolivia plus the Chaco region (one province); (b) a sample of households that live near the schools in the treatment or control group for education projects; and (c) a sample of households that will benefit from water or latrine projects.

To analyze how well SIF investments are actually targeted to the poor, the study uses the baseline (pre-SIF investment) data and information on where SIF investments were later placed to calculate the probability that individuals will be SIF beneficiaries conditional on their income level. The study then combines the baseline and follow-up survey data to estimate the average impact of SIF in those communities that received a SIF investment, using regression techniques. In addition to average impact, it explores whether the characteristics of communities, schools, or health centers associated with significantly greater than average impacts can be identified.

In education, for which SIF investments were randomly assigned among a larger pool of equally eligible communities, the study applies the “ideal” randomized experiment design (in which the counterfactual can be directly observed). In health and sanitation projects, in which pro-

jects were not assigned randomly, the study uses the “instrumental variable” method to compensate for the lack of a direct counterfactual. Instrumental variables are correlated with the intervention but do not have a direct correlation with the outcome.

IV. Results

SIF II investments in education and health do result in a clear improvement in infrastructure and equipment. Education projects have little impact on school dropout rates, but school achievement test scores among sixth graders are significantly higher in SIF schools. In health, SIF investments raise health service utilization rates and reduce mortality. SIF water projects are associated with little improvement in water quality but do improve water access and quantity and also reduce mortality rates.

A comparison of the randomized versus matched-comparison results in education shows that the matched-comparison approach yields less comparable treatment and comparison groups and therefore less robust results in discerning program impact. In illustration of this finding, evidence of improvements in school infrastructure (which one would clearly expect to be present in SIF schools) is picked up in the randomized evaluation design but not in the matched-comparison design.

Finally, the results show that SIF II investments are generally not well targeted to the poor. Health and sanitation projects benefit households that are relatively better off in terms of per capita income, and there is no relationship between per capita income and SIF education benefits.

V. Policy Application

The results on targeting reveal an inherent conflict between the goal of targeting the poor and the demand-driven nature of SIF. With the introduction of the popular participation law in 1994, subprojects had to be submitted through municipal governments. The targeting results suggest that even in a highly decentralized system it is important to monitor targeting processes. In the Bolivian case, it appears that better-off, more organized communities, rather than the poorest, are those most likely to obtain SIF investments. In the case of SIF sanitation projects in particular, the bias against poorest communities may be hard to correct. Investment in basic sanitation is most efficient in populated areas that already have access to a water system so that the project can take advantage of economies of scale.

The fact that SIF investments have had no perceptible impact on school attendance has prompted a restructuring of SIF interventions in this sec-

tor. Rather than focusing solely on providing infrastructure, projects will provide a combination of inputs designed to enhance school quality. Similarly, disappointing results on water quality (which shows no improvement resulting from SIF projects compared with the preexisting source) have generated much attention, and project design in this sector is being rethought.

VI. Lessons Learned

Effectiveness of the Randomization Technique. The randomized research design, in which a control group is selected at random from among potential program beneficiaries, is far more effective at detecting program impact than the matched-comparison method of generating a control group. Randomization must be built into program design from the outset in determining the process through which program beneficiaries will be selected, and random selection is not always feasible. However, when program funds are insufficient to cover all beneficiaries, an argument can be made for random selection from among a larger pool of qualified beneficiaries.

Importance of Institutionalizing the Evaluation Process. Evaluations can be extremely complex and time consuming. The Bolivia evaluation was carried out over the course of seven years in an attempt to rigorously capture project impact, and achieved important results in this regard. However, the evaluation was difficult to manage over this length of time and given the range of different actors involved (government agencies and financing institutions). Management and implementation of an evaluation effort can be streamlined by incorporating these processes into the normal course of local ministerial activities from the beginning. Further, extensive evaluation efforts may be best limited to only a few programs—for example, large programs in which there is extensive uncertainty regarding results—in which payoffs of the evaluation effort are likely to be greatest.

VII. Evaluation Costs and Administration

Costs. The total estimated cost of the Bolivia SIF evaluation to date is \$878,000, which represents 0.5 percent of total project cost. Data collection represents a relatively high proportion of these costs (69 percent), with the rest being spent on travel, World Bank staff time, and consultants.

Administration. The evaluation was designed by World Bank staff and financed jointly by the World Bank, KfW, and the Dutch, Swedish, and Danish governments. Survey work was conducted by the Bolivian

National Statistical Institute and managed by SIF counterparts for the first round and later the Ministry of Finance for the second round.

VIII. Source

Pradhan, Menno, Laura Rawlings, and Geert Ridder. 1998. "The Bolivian Social Investment Fund: An Analysis of Baseline Data for Impact Evaluation." *World Bank Economic Review* 12 (3): 457–82.

Annex 1.5: Impact of Active Labor Programs: Czech Republic

I. Introduction

Project Description. Many developing countries face the problem of retraining workers when state-owned enterprises are downsized. This is particularly complicated in transition economies that are also characterized by high unemployment and stagnant or declining wages. However, all retraining programs are not created equal. Some are simply disguised severance pay for displaced workers; others are disguised unemployment programs. This makes the case for evaluation of such programs particularly compelling.

Training programs are particularly difficult to evaluate, however, and the Czech evaluation is no exception. Typically, several different programs are instituted to serve different constituencies. There are also many ways of measuring outcomes, including employment, self-employment, monthly earnings, and hourly earnings. More than with other types of evaluations, the magnitude of the impact can be quite time-dependent: very different results can be obtained depending on whether the evaluation is one month, six months, one year, or five years after the intervention.

Highlights of Evaluation. This evaluation quantified the impact of four active labor market programs (ALP) in the Czech Republic using quasi-experimental design methods—matching ALP participants with a similar group of nonparticipants. Both administrative and follow-up survey data were used in an ex post evaluation of a variety of different outcomes: duration of unemployment, likelihood of employment, self-employment, and earnings. Regression analysis is used to estimate the impact of each of the five programs on these outcomes, controlling for baseline demographic characteristics.

Several important lessons were learned from this evaluation. One set of lessons is practical: how to design quite a complex evaluation, how to use administrative data, how to address the problems associated with administering the survey, and the mechanics of creating the matched sample. The second is how to structure an analysis to provide policy-relevant information—made possible by a detailed evaluation of the impact by subgroup. This led to a policy recommendation to target ALP programs to particular types of clients and concluded that one type of ALP is not at all effective in changing either employment or earnings.

II. Research Questions and Evaluation Design

This is part of a broader evaluation of four countries: the Czech Republic, Poland, Hungary, and Turkey. The common context is that each country had high unemployment, partially caused by the downsizing of state-owned enterprises, which had been addressed with passive income support programs, such as unemployment benefits and social assistance. This was combined with the ALPs that are the subject of this evaluation. The five ALPs are Socially Purposeful Jobs (new job creation), Publicly Useful Jobs (short-term public employment), Programs for School Leavers (subsidies for the hiring of recent graduates), Retraining (occupation-specific training lasting a few weeks to several months), and Programs for the Disabled and Disadvantaged. The last is rather small and not included in the evaluation.

There are two research questions. One is to examine whether participants in different ALPs are more successful at reentering the labor market than are nonparticipants and whether this varies across subgroups and with labor market conditions. The second is to determine the cost-effectiveness of each ALP and make suggestions for improvement.

The evaluation is an *ex post*, quasi-experimental design—essentially a matched cohort. The participant group is matched with a constructed nonparticipant group (with information drawn from administrative records) on people who registered with the state employment service but were not selected for the ALP. The fundamental notion is that an individual is selected at random from the ALP participant group. This individual's outcomes are then compared with those for individuals in the nonparticipant group (based on age, gender, education, number of months unemployed, town size, marital status, and last employment type). The evaluation is particularly strong in its detailed analysis of the comparison versus the participant group.

There are inevitably some problems with this approach, and they have been extensively addressed elsewhere (Burtless 1995, and Heckman and Smith 1995). One obvious concern that is endemic to any nonrandomized trial is that participants may have been “creamed” by the training program on the basis of characteristics unobservable to or unmeasured by the researchers. The second major concern is that nonparticipants may have substituted other types of training for public training in the case of the retraining program. The third concern is that subsidies to employ workers may have simply led to the substitution of one set of workers by another.

III. Data

One very interesting component of this evaluation was the use of government administrative data to create the sample frame for the survey.

The team thus visited the Ministry of Labor and Social Affairs (MOLSA) in Prague and three local labor market offices to develop an understanding of both the administration and implementation of ALPs and of the administrative data on ALP participants. From this, 20 districts were chosen for survey, based on criteria of geographic dispersion and variation in industrial characteristics. There was also a broad range of unemployment rates across districts. The survey contained both quantitative questions about the key program outcomes and qualitative questions about the participants' rating of the program.

Another valuable component was the implementation of a pilot survey in four districts. This approach, which is always important, identified not only technical problems but also a legal problem that can often arise with the use of administrative records. This issue is the interpretation of privacy law: in this case, MOLSA did not permit a direct mailing but required that potential respondents give permission to the labor office to allow their addresses to be given out. This delayed the evaluation schedule, increased costs, and dramatically lowered the response rate.

The survey was conducted in early 1997 on a random sample of 24,973 labor office registrants who were contacted. Of these, 9,477 participated in ALP during 1994–95. The response rate for nonparticipants was 14 percent; for participants it was 24.7 percent, resulting in a total number of 4,537 respondents. The dismal response rate was directly attributable to the legal ruling: most people did not respond to the initial request, but among those who did allow their address to be given, the response rate was high. Worse, the resulting bias is unknown.

IV. Econometric Techniques

The difficulty of measuring both the temporal nature and the complexity of labor market outcomes is illustrated by the use of seven different outcome measures: percent currently employed, percent currently self-employed, percent ever employed, length of unemployment, length of receiving unemployment payments, total unemployment payments, and current monthly earnings

The evaluation approach, however, was fairly straightforward in its use of both simple differences across groups and ordinary least squares with group-specific dummies to gauge the impact of the interventions. The overall impact was calculated, followed by estimated impacts by each of the subgroup categories (age, sex, education, and, for earnings outcomes, size of firm). This last analysis was particularly useful because it identified subgroups of individuals for whom, in fact, the impact of the interventions was different, leading to quite different policy implications. Indeed, a major recommendation of the evaluation was that the ALPs be more tightly targeted.

V. Who Carried It Out

The evaluation was part of a four-country cross-country evaluation of active labor programs, with the express motivation of understanding the impact of ALPs under different economic conditions. The evaluation was supervised by a project steering committee, which had representatives from the World Bank, from each of the four countries, from the external financing agencies, and from the technical assistance contractors (Abt Associates and the Upjohn Institute).

The team contracted with a private survey firm to carry out the survey itself—for data quality reasons as well as to reduce the possibility of intimidation if the local labor office were to carry out the survey. It is worth making the point that the credibility of the study could be contaminated if the employment service were responsible for conducting the survey. Indeed, this moral hazard problem is generally an important one if the agency responsible for training is also responsible for collecting information on the outcomes of that training.

VI. Results

The results are typical of evaluations for training programs. Some interventions appear to have some (albeit relatively weak) impacts for some types of workers in some situations. A strong point of the evaluation is that it does identify one program that appears to have wasted money—no impact was shown either overall or for any subgroup. Another strong point is the presentation of the evaluation itself, which is particularly important if the evaluation is to be read by policymakers. Here, tables are provided for each program summarizing the combined benefits in terms of wages and employment, both in aggregate and for each subgroup.

A very negative point is that, despite the initial promise, no cost-benefit analysis was performed. It would have been extremely useful to have the summary benefit information contrasted with the combined explicit and implicit cost of the program. Thus, although, for example, the evaluators found that one program increased the probability of employment across the board, it should be noted that this came at a cost of a nine-month training program. A full calculation of the rate of return of investment would have combined the explicit cost of the program with the opportunity cost of participant time and compared this with the increase in earnings and employment.

VII. Lessons Learned

Several important lessons were learned from this study. First among these are the pragmatic components discussed in the introduction, par-

ticularly the importance of taking the political environment into consideration in designing an evaluation scheme. The inability to convince the employment service of the importance of the evaluation project meant that the survey instrument was severely compromised. Second, the study provides a useful demonstration of the construction of a matched sample. Finally, the evaluation provides a good illustration of the importance of conducting analysis not just in aggregate but also on subgroups, with the resultant possibility of fruitful targeted interventions.

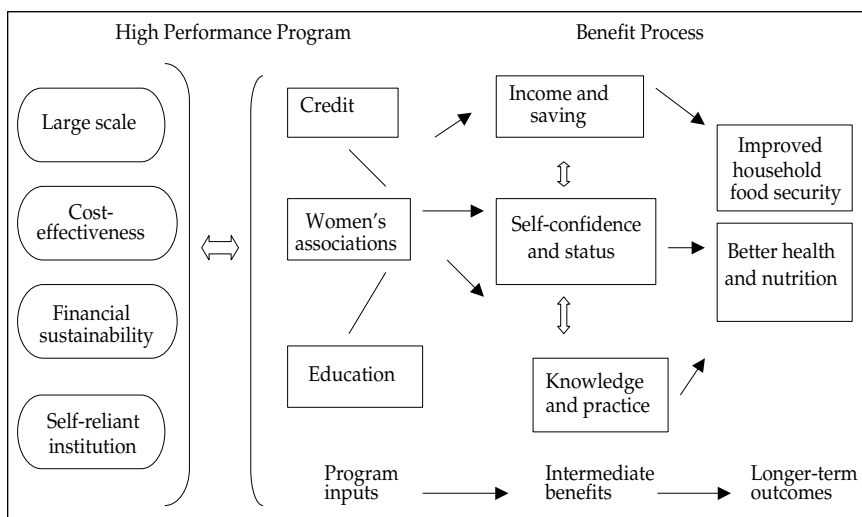
VIII. Sources

Benus, Jacob, Grover Neelima, Jiri Berkovsky, and Jan Rehak. 1998. *Czech Republic: Impact of Active Labor Market Programs*. Cambridge, Mass., and Bethesda, Md.: Abt Associates, May.

Burtless, Gary. 1995. "The Case for Randomized Field Trials in Economic and Policy Research." *Journal of Economic Perspectives* 9 (2): 63–84.

Heckman, James J., and Jeffrey A. Smith. 1995. "Assessing the Case for Social Experiments." *Journal of Economic Perspectives* 9 (2) : 85–110.

Schematic Used for Designing the Czech Active Labor Programs Evaluation



Annex 1.6: Impact of Credit with Education on Mothers' and Their Young Children's Nutrition: Lower Pra Rural Bank Program in Ghana

I. Introduction

Project Description. The Credit with Education program combines elements of the Grameen Bank program with education on the basics of health, nutrition, birth timing and spacing, and small business skills. The aim is to improve the nutritional status and food security of poor households in Ghana. Freedom from Hunger, together with the Program in International Nutrition at the University of California Davis, provided Credit with Education services to poor rural women in the Shama Ahanta East District of the Western Region of Ghana. A partnership was formed with five rural banks to deliver such services—more than 9,000 loans, totaling \$600,000, were made by March 1997 with a repayment rate never below 92 percent.

Highlights of Evaluation. The evaluation is interesting for three reasons. First, the sample design was quite appropriate: the program was administered to 19 communities and data were collected on three different sample groups of women: those who participated at least one year, those who did not participate but were in the program communities, and those in control communities. Second, the study had a clear description of its underlying approach: it identified and evaluated both intermediate and longer-term outcomes. Finally, it provided a nice blend of both qualitative and quantitative results, often highlighting the quantitative outcomes with an anecdotal illustrative example.

II. Research Questions and Evaluation Design

The research questions focused on the program's effects on (a) the nutritional status of children; (b) women's economic capacity (income, savings, time) to invest in food and health care; (c) women's knowledge and adoption of breastfeeding, weaning, and diarrhea management and prevention practices; and (d) women's ability to offer a healthy diet to their children.

In doing this, the evaluation separated out the ultimate goals of improved household food security and nutritional status from the intermediate benefits of changing behavior, reducing poverty, and female empowerment.

A quasi-experimental design was used in fielding two surveys (in 1993 and 1996) to evaluate the impact of the strategy on children's nutritional status; mothers' economic capacity, women's empowerment, and mothers' adoption of child health and nutrition practices. A total of 299 mother-and-child pairs were surveyed in the first period and 290 different pairs in the second period. Both qualitative and quantitative information was gathered.

The evaluation design was quite complex. The Lower Pra Rural Bank identified 19 communities that had not yet had Credit with Education services, and the consultants divided communities into large and small (greater or less than 800) and then again by whether they were close to a main road. Within each stratification, the 13 of the 19 communities were assigned either to a treatment or to a control group. Three were given the treatment for political reasons and three communities were selected as matched controls to the politically selected three based on their proximity, commercial development, size, and access to main roads. Two communities dropped out because of lack of interest and the small number of communities in the classification. Thus, in the follow-up study only 17 communities were surveyed.

Ten mother-and-child pairs, with children aged 12 to 23 months, were chosen for the baseline surveys from small communities, and 30 from the large communities. Two important problems arose as a result. The first is that this construction did not allow the surveys to follow the same women over time because few women in the baseline survey also had infants in the 1996 survey. The second problem was that the age restriction cut the second sample so much that it was extended to women with children under three years of age in 1996. A major advantage of this complex evaluation design was that it was possible to classify women in the baseline samples as future participants and future nonparticipants.

Three types of women were surveyed: participants, nonparticipants in the program communities, and residents in control communities. All participants were included; the latter two types were randomly selected from women with children under three. It is worth noting that the total sample size (of 360) was calculated based on the standard deviations found in previous studies, a requirement that the sample be able to detect a 0.4 difference in the z-score values of the control and target groups and with a target significance level of 0.05 and a power of 0.8.

III. Data

Both quantitative and qualitative data were collected on the household, mother and child, focusing on both intermediate and long-term measures—and particularly the multidimensional nature of the outcomes.

For the intermediate outcomes, this led to a set of questions attempting to measure women's economic capacity (incomes, profit, contribution to total household income, savings, entrepreneurial skill, and expenditures on food and households). Similarly, another set of measures addressed the woman's knowledge of health and nutrition (breastfeeding, child feeding, diarrhea treatment and prevention, and immunization). Yet another set captured women's empowerment (self-confidence and hope about the future, status and decisionmaking in the household, and status and social networks in the community). For the ultimate outcomes, such as nutritional status and food security, more direct measures were used (anthropometric measures for the former, questions about hunger in the latter case).

Although a total sample size of 360 mother-and-child pairs was planned, only 299 pairs were interviewed in the first survey (primarily because two communities were dropped) and 290 in the second. Mother and household characteristics were compared across each of the three groups and no significant differences were found.

IV. Econometric Techniques

The econometric techniques used are fairly straightforward and exploited the strength of the survey design. The group mean is calculated for each of the varied outcome measures used, and then t-tests are performed to examine whether differences between controls and participants are significant. This is essentially a simple-difference approach. These are well supplemented with graphics.

A series of major questions were not addressed, however. First, the sample design was clustered—and because, almost by construction, the outcomes of each individual mother-and-child pair will be correlated with the others in the community, the standard errors will be biased down and the t-statistics spuriously will be biased up. In the extreme case, in which all the individual outcomes are perfectly correlated with each other, the sample size is actually 17 rather than 300. This will lend significance to results that may, in fact, not be significant. Second, although the design was explicitly stratified, the impact of that stratification was not addressed: either whether large or small communities benefited more or whether communities close to a road were better off than those a long way away from a road. This is particularly surprising, since presumably the reason to have such a sample design is to examine the policy implications. Third, although selection bias problems are discussed, there is no formal analysis of or correction for this fundamental problem. Finally, although there were significant differences in item non-response rates, which suggests the potential for selection bias even within the survey, this was neither addressed nor discussed.

V. Who Carried It Out

An international not-for-profit institute, Freedom from Hunger, developed the Credit with Education program and collaborated with the Program in International Nutrition at the University of California Davis, in evaluating it. The institute partnered with the Lower Pra Rural Bank (an autonomous bank, regulated by the Bank of Ghana), and subsequently four other rural banks in Ghana, to deliver the program. The Lower Pra Rural Bank played a role in identifying and selecting the communities to be surveyed.

VI. Results

The intermediate goals were generally achieved: although women's incomes and expenditures did not increase, women's entrepreneurial skills and savings were significantly higher. Women's health and nutrition knowledge was generally improved. Women were also more likely to feel empowered. In terms of the ultimate goals the evaluation suggested that the program did improve household food security and child nutritional status but not maternal nutritional status.

VII. Lessons Learned

A key contribution of the evaluation is the very interesting sample design: the stratification and the choice of participant and nonparticipant groups with respect to their future participation is a very useful approach. Another lesson is the productive use of many outcome dimensions—sometimes on quite nonquantitative factors such as women's empowerment. The other key lesson is the value of nonquantitative data to illustrate the validity of quantitative inferences.

VIII. Source

MkNelly, Barbara, and Christopher Dunford (in collaboration with the Program in International Nutrition, University of California Davis). 1998. "Impact of Credit with Education on Mothers' and their Young Children's Nutrition: Lower Pra Rural Bank Credit with Education Program in Ghana." Freedom from Hunger Research Paper No. 4, March.

Annex 1.7: Textbooks and Test Scores: Evidence from a Prospective Evaluation in Kenya

I. Introduction

Project Description. Evaluating the effect of different types of education expenditure on student outcomes is particularly important in developing countries. Prior studies have suggested that the provision of textbooks is a cost-effective way of increasing test scores, and Kenya, with the extraordinarily scarce resources available to educators, makes a good case study. The evaluators note that only one in six children in grades 3, 4, and 5 has textbooks; this rises to one in four in later grades. In addition, physical facilities are extremely poor with many children sitting on the floor to learn.

The evaluation assessed the impact on learning outcomes of a 1996 program in which all grades in a randomly selected subset of 25 out of 100 rural Kenyan primary schools were provided with textbooks. English textbooks were given to grades 3 through 7, with a ratio of 6 textbooks to every 10 children; mathematics textbooks to grades 3, 5, and 7, with a 50 percent ratio; and science textbooks to grade 8, with a 60 percent ratio. In addition, each class was provided with a teacher's guide. Achievement tests were given to the students before textbooks were distributed and then again 10 months later. The same tests were also given to the control schools. This approach combines a randomized design with reflexive comparisons.

Highlights of Evaluation. This evaluation is an excellent illustration of developing and implementing a good survey design and then following that up with appropriate econometric techniques. It is particularly strong in showing how to draw inferences on level outcomes with stacked data, the use of difference-in-difference estimators, how to address selection and attrition bias, as well as measurement error and crowding-out issues. Another very interesting component of the evaluation is the focus on the intervention's impact on students in all parts of the distribution. Finally, the recognition and analysis of potential secondary effects is a very good example of looking at all dimensions of an intervention.

II. Research Questions and Evaluation Design

The main focus of the research is to evaluate the effect of textbooks on learning outcomes. Because this is a complex concept, the outcomes are measured as the difference between textbook and comparison schools in

several dimensions: posttest scores, test score gains, differences between subject-grade combinations that did and did not receive textbooks, and child and teacher activity. The evaluation also considered other (often ignored) secondary effects, particularly the possibility that the provision of such a subsidy would reduce parental involvement, particularly in terms of crowding out other fundraising.

The evaluation design is quite complex. The Ministry of Education chose 100 needy schools for the intervention in 1995. These were divided into four groups—first on the basis of geography, then on an alphabetical basis within the geography. There was then an ordered assignment, on the basis of the alphabet, of each school to each of the four groups. Textbook assistance was staggered to go to the first group in 1996, the second group in 1997, and so on. Mathematics, English, and science textbooks were provided to different grades—primarily grades 3 through 7.

III. Data

Math, English, and science exams were given to children in all these grades in each of the 100 schools before textbooks were distributed. The evaluation itself, however, makes use of pretests that were administered in grades 3 through 7 in October 1996 and posttests in October 1997. There are therefore data on some 8,800 students (in all grades) for each subject in the 100 schools and a total of over 26,000 observations. Because 25 schools received the textbooks in this period, students in these schools become the “textbook” group; the other 75 are the comparison group. In addition to test scores, data were also collected on school finances and on pedagogical methods.

Information on classroom utilization of textbooks was gathered by trained observers who visited each school and took minute-by-minute notes on eight possible classroom activities (ranging from general teacher and pupil activity to the use of textbooks by teachers and pupils). These notes covered 15 minutes and were then used to construct percentages of time spent by teachers and students in each different activity for a total of 551 class periods. Four to five students in each class were interviewed by field staff, who filled out a questionnaire on the basis of their responses.

Finally, data were gathered on school finances from a 1997 school and school committee questionnaire, which asked about fund-raising activities.

IV. Econometric Techniques

It is worth noting the interesting issues generated by this sampling technique. Test scores within a school are likely to be correlated with each other, as are within-class scores. Similarly, test scores for different subjects

taken by the same child will be correlated. The intervention can also be evaluated in terms of the impact on outcomes on student learning levels or on student learning gains. In general, the effect of an intervention should be robust to different econometric techniques and different ways of looking at the data, and this was certainly the case here.

The evaluation proceeds by first providing estimates from a simple dummy-variable-level regression, with treatment dummies for each grade-subject combination with school, grade, and subject random effects (the dependent variable is the change in test scores from the pre- to the posttest). One attractive feature of this is that the dummies can be combined in very useful ways:

- Pooling several grades to estimate the impact of textbooks for a subject
- Pooling all test scores to estimate the average impact of textbooks for a grade; and
- Pooling all grades and subjects to estimate the weighted average impact of textbooks for all grades and subjects.

Clearly, the structure of the random effects varies with each approach, and the evaluation is very clear in this component.

The evaluation then proceeds with a difference-in-difference approach, which is relatively straightforward in that it simply compares post- and pretest scores between control and treatment schools.

The third approach, which is a little more complicated because it exploits within-school variation, deserves discussion. The regression applied here involves regressing test scores on dummies that capture whether the students were (a) in a textbook school and (b) in a subject-grade combination that received a textbook. This reduces problems introduced by school heterogeneity as well as sample selection problems—in the latter case because it captures the effect on test scores for the same student depending on whether or not the student received a textbook. It does assume, however, that test scores in different grade-subject combinations can be added and subtracted, and this very strong assumption may be the reason for very different results from this approach.

A recurring theme in evaluations is the desire to capture not just the average effect of the intervention but also the effect on subgroups of recipients. This evaluation provides a very useful illustration of the use of interaction terms and quantile regression. The former approach involves interaction between initial test scores and textbook dummies to capture the effect of textbooks on better versus poorer students, using both actual and instrumented values (initial test scores are correlated with the error term, causing a bias). The second approach, which involves using quantile regression, is also useful and increasingly popular. More specif-

ically, since least squares regression only captures the average impact of the textbook program, quintile regressions allow the effect of the treatment to vary depending on where the student is in the distribution.

The evaluation is also particularly strong in providing an application of how to look for selection and attrition bias. The major potential source of problems in this intervention is differential promotion and repetition rates between textbook and comparison schools. For example, children might be differentially promoted from grade 2 (a nontextbook grade) to grade 3 (a textbook grade) in textbook schools. Differential promotion biases down the results in the classes that the worst students are added to, and possibly biases up the results in the classes they came from. These two effects were captured in the evaluation by reestimating the model in two ways: dropping all repeaters from both sets of schools and dropping the worst students in each grade. The robustness of the results under both approaches confirmed the impact of the intervention.

Finally, in an illustration of considering the importance of secondary effects, the evaluation quantified the impact of textbook provision on parent fundraising. They found that the intervention did crowd out parent contributions—the amount of non-ICS aid received by comparison schools was \$465 and for textbook schools \$267 (the average value of ICS textbooks was \$485). They used simple regression analysis and also investigated, and confirmed, the hypothesis that smaller schools had more crowding out than larger schools.

Who Carried It Out. A Dutch nonprofit organization, International Christelijk Steunfonds, funded the project. The evaluation was performed by a Massachusetts Institute of Technology professor (Kremer) and two World Bank economists (Paul Glewwe and Sylvie Moulin). Some of the costs were covered by the National Science Foundation and the World Bank research committee.

V. Results

The result of this evaluation was in marked contrast to the results of other evaluations of textbook interventions. The basic result was that there was no significant impact of textbooks on learning outcomes on average, but that there was a significant effect for better students. This was robust to different estimation techniques and cuts of the data.

VI. Lessons Learned

The most useful lesson learned from this evaluation was the importance of using different econometric techniques to check for the robustness of

the empirical results. Even though the data collection was close to ideal, it is important that the estimated impact of the intervention remain roughly the same with different econometric assumptions and model specifications. The application of quantile regression and interaction terms was also a very useful way to analyze the impact on different subgroups of the population. Finally, it is important to look for and identify secondary effects—in this case, the potential for crowding out.

VIII. Source

Glewwe, Paul, Michael Kremer, and Sylvie Moulin. 1998. "Textbooks and Test Scores: Evidence from a Prospective Evaluation in Kenya." Development Research Group (DECRG), World Bank, Washington, D.C. Processed.

Annex 1.8: Evaluating Kenya's Agricultural Extension Project

I. Introduction

Project Description. The first National Extension Project (NEP-I) in Kenya introduced the Training and Visit (T&V) system of management for agricultural extension services in 1983. The project had the dual objectives of institutional development and delivering extension services to farmers with the goal of raising agricultural productivity. NEP-II followed in 1991 and aimed to consolidate the gains made under NEP-I by increasing direct contact with farmers, improving the relevance of extension information and technologies, upgrading skills of staff and farmers, and enhancing institutional development.

Impact Evaluation. The performance of the Kenyan extension system has been controversial and is part of the larger debate on the cost-effectiveness of the T&V approach to extension. Despite the intensity of the debate, the important role of agricultural extension services in the World Bank's development strategy for Africa, and the large volume of investments made, very few rigorous attempts have been made to measure the impact of T&V extension. In the Kenyan case, the debate has been elevated by very high estimated returns to T&V reported in an earlier study, and the lack of convincingly visible results, including the poor performance of Kenyan agriculture in recent years.

The disagreement (between the Operations Evaluation Department and the Africa Region of the World Bank) over the performance of NEP-I has persisted pending this evaluation, which takes a rigorous empirical approach to assess the program's impact on agricultural performance. Using the results-based management framework, the evaluation examines the impact of project services on farm productivity and efficiency. It also develops measures of program outcomes (that is, farmer awareness and adoption of new techniques) and outputs (for example, frequency and quality of contact) to assess the performance of the extension system and to confirm the actual, or the potential, impact.

II. Evaluation Design

The evaluation strategy illustrates best-practice techniques in using a broad array of evaluation methods in order to assess program implementation, output, and its impact on farm productivity and efficiency.

(No attempt is made to study the impact on household welfare, which is likely to be affected by a number of factors far beyond the scope of T&V activities.) It draws on both quantitative and qualitative methods so that rigorous empirical findings on program impact could be complemented with beneficiary assessments and staff interviews that highlight practical issues in the implementation process. The study also applied the contingent valuation method to elicit farmers' willingness to pay for extension services. [The contingent valuation method elicits individuals' use and nonuse values for a variety of public and private goods and services. Interviewees are asked to state their willingness to pay (accept) to avoid (accept) a hypothetical change in the provision of the goods or services—that is, the “contingent” outcome. In this case, farmers were asked how much they would be willing to pay for continued agricultural extension services should the government cease to provide them.]

The quantitative assessment is complicated by the fact that the T&V system was introduced on a national scale, preventing a with-program and without-program (control group) comparison. The evaluation methodology therefore sought to exploit the available preproject household agricultural production data for limited before-and-after comparisons using panel data methods. For this, existing household data were complemented by a fresh survey to form a panel. Beneficiary assessments designed for this study could not be conducted, but the evaluation draws on the relevant findings of two recent beneficiary assessments in Kenya. The study is noteworthy in that it draws on a range of preexisting data sources in Kenya (household surveys, participatory assessments, and so forth), complemented by a more comprehensive data collection effort for the purpose of the evaluation.

III. Data Collection and Analysis Techniques

The evaluation approach draws on several existing qualitative and quantitative data sources. The quantitative evaluation is based largely on a 1998 household survey conducted by the World Bank's Operations Evaluation Department. This survey generates panel data by revisiting as many households as could be located from a 1990 household survey conducted by the Africa Technical Department, which in turn drew from a subsample of the 1982 Rural Household Budget Survey. (These three surveys generate a panel data set for approximately 300 households. The surveys cover household demographics, farm characteristics, and input-output data on agricultural production; the 1990 and 1998 surveys also collect information on contact with extension services, including awareness and adoption of extension messages.) These data are supplemented by a survey of the extension staff, several recent reviews of the extension

service conducted or commissioned by the Ministry of Agriculture, and individual and focus group discussions with extension staff. The study also draws on two recent beneficiary assessments: a 1997 study by ActionAid Kenya, which elicited the views of users and potential users of Kenya's extension services; and a 1994 Participatory Poverty Assessment, which inquired about public services, including extension, and was carried out jointly by the World Bank, British Overseas Development Administration, African Medical and Research Foundation, UNICEF, and the government of Kenya.

The analysis evaluates both the implementation process and the outcome of the Kenyan T&V program. The study evaluates institutional development by drawing on secondary and qualitative data—staff surveys, interviews, and the ministry's own reviews of the extension service. Quality and quantity of services delivered are assessed by using a combination of the findings of participatory (beneficiary) assessments, staff surveys, and through measures of outreach and the nature and frequency of contact between extension agents and farmers drawn from the 1998 OED survey. The survey data are also used to measure program outcomes, measured in terms of farmer awareness and adoption of extension recommendations.

The program's results—its actual impact on agricultural production in Kenya—are evaluated by relating the supply of extension services to changes in productivity and efficiency at the farm level. Drawing on the household panel data, these impacts are estimated by using the data envelopment analysis, a nonparametric technique, to measure changes in farmer efficiency and productivity over time, along with econometric analysis measuring the impact of the supply of extension services on farm production. Contingent valuation methods are used to directly elicit the farmers' willingness to pay for extension services.

IV. Results

The institutional development of NEP-I and NEP-II has been limited. After 15 years, the effectiveness of extension services has improved little. Although there has been healthy rethinking of extension approaches recently, overall the extension program has lacked the strategic vision for future development. Management of the system continues to be weak, and information systems are virtually nonexistent. The quality and quantity of service provision are poor. Beneficiaries and extension service staff alike report that visits are infrequent and ineffective. Although there continues to be unmet demand for technically useful services, the focus of the public extension service has remained on simple and basic agronomic messages. Yet the approach taken—a high intensity of contact

with a limited number of farmers—is suited to deliver more technical information. The result has been a costly and inefficient service delivery system. Extension activities have had little influence on the evolution of patterns of awareness and adoption of recommendations, which indicates limited potential for impact. In terms of the actual impact on agricultural production and efficiency, the data indicate a small positive impact of extension services on technical efficiency but no effect on allocative or overall economic efficiency. Furthermore, no significant impact of the supply of extension services on productivity at the farm level could be established by using the data in hand. The data do show, however, that the impact has been relatively greater in the previously less productive areas, where the knowledge gap is likely to have been the greatest. These findings are consistent with the contingent valuation findings. A vast majority of farmers, among both the current recipients and nonrecipients, are willing to pay for advice, indicating an unmet demand. However, the perceived value of the service, in terms of the amount offered, is well below what the government is currently spending on delivering it.

V. Policy Implications

The Kenya Extension Service Evaluation stands out in terms of the array of practical policy conclusions that can be derived from its results, many of which are relevant to the design of future agricultural extension projects. First, the evaluation reveals a need to enhance targeting of extension services, focusing on areas and groups in which the difference between the average and best practice is the greatest and hence the impact is likely to be greatest. Furthermore, advice needs to be carefully tailored to meet farmer demands, taking into account variations in local technological and economic conditions. Successfully achieving this level of service targeting calls for regular and timely flows of appropriate and reliable information, and the need for a monitoring and evaluation system to provide regular feedback from beneficiaries on service content.

To raise program efficiency, a leaner and less-intense presence of extension agents with wider coverage is likely to be more cost-effective. There are not enough technical innovations to warrant a high frequency of visits, and those currently without access demand extension services. The program's blanket approach to service delivery, relying predominantly on a single methodology (farm visits) to deliver standard simple messages, also limits program efficiency. Radio programs are now popular, younger farmers are more educated, and alternative providers (non-governmental organizations) are beginning to emerge in rural Kenya. A flexible pluralistic approach to service delivery, particularly one that uses

lower-cost means of communication, is likely to enhance the program's cost-effectiveness.

Finally, the main findings point to the need for institutional reform. As with other services, greater effectiveness in the delivery of extension services could be achieved with more appropriate institutional arrangements. The central focus of the institution should be the client (farmer). Decentralization of program design, including participatory mechanisms that give voice to the farmer (such as cost sharing and farmer organizations) should become an integral part of the delivery mechanism. Financial sustainability is critical. The size and intensity of the service should be based on existing technological and knowledge gaps and the pace of flow of new technology. Cost recovery, even if only partial, offers several advantages: it provides appropriate incentives, addresses issues of accountability and quality control, makes the service more demand-driven and responsive, and provides some budgetary respite. Such decentralized institutional arrangements remain unexplored in Kenya and in many extension programs in Africa and around the world.

VI. Evaluation Costs and Administration

Costs. The total budget allocated for the evaluation was \$250,000, which covered household survey data collection and processing (\$65,000—though this probably is an underestimate of actual costs); extension staff survey, data, and consultant report (\$12,500); other data collection costs (\$12,500); and a research analyst (\$8,000). Approximately \$100,000 (not reflected in the official costs) of staff costs for data processing, analysis, and report writing should be added to fully reflect the study's cost.

Administration. To maintain objectivity and dissociate survey work from both the government extension service and the World Bank, the household survey was implemented by the Tegemeo Institute of Egerton University, an independent research institute in Kenya. The analysis was carried out by Madhur Gautam of the World Bank.

VII. Lessons Learned

- The combination of theory-based evaluation and a results-based framework can provide a sound basis for evaluating the impact of project interventions, especially when many factors are likely to affect intended outcomes. The design of this evaluation provided for the measurement of key indicators at critical stages of the project cycle, linking project inputs to the expected results to gather sufficient evidence of impact.

- An empirical evaluation demands constant and intense supervision. An evaluation can be significantly simplified with a well-functioning and high quality monitoring and evaluation system, especially with good baseline data. Adequate resources for these activities are rarely made available. This evaluation also benefited tremendously from having access to some, albeit limited, data for the preproject stage and also independent sources of data for comparative purposes.
- Cross-validation of conclusions using different analytical approaches and data sources is important to gather a credible body of evidence. Imperfect data and implementation problems place limits on the degree of confidence that individual methods can provide answers to key evaluative questions. Qualitative and quantitative assessments strongly complement each other. The experience from this evaluation indicates that even in the absence of participatory beneficiary assessments, appropriately designed questions can be included in a survey to collect qualitative as well as quantitative information. Such information can provide useful insights to complement quantitative assessments.
- If properly applied, contingent valuation can be a useful tool, especially in evaluating the value of an existing public service. The results of the application in this evaluation are encouraging, and the responses appear to be rational and reasonable.

VIII. Sources

World Bank. 1999. *World Bank Agricultural Extension Projects in Kenya: An Impact Evaluation*. Operations Evaluation Department, Report no. 19523. Washington, D.C.

In addition, the following working papers are also available from the World Bank Operations Evaluation Department:

The Efficacy of the T&V system of Agricultural Extension in Kenya: Results from a Household Survey

Awareness and Adoption of Extension Messages

Reconsidering the Evidence on Returns to T&V Extension in Kenya

Farmer Efficiency and Productivity Change in Kenya: An Application of the Data Envelopment Analysis

The Willingness to Pay for Extension Services in Kenya: An Application of the Contingent Valuation Method

Annex 1.9: The Impact of Mexico's Retraining Program on Employment and Wages (PROBECAT)

I. Introduction

This case is somewhat unusual in that three evaluations of the program have been carried out—first, by the World Bank using data from 1992 (Revenge, Riboud, and Tan 1994); second, by the Mexican Ministry of labor using data from 1994 (STPS 1995); and third, an update by the World Bank (Wodon and Minowa 1999). The methodologies used for the first two evaluations were quite similar, and they gave similar results. Methodological enhancements in the third evaluation led to fairly different findings and policy conclusions. The fact that the results differ substantially between the first two evaluations and the third highlights the importance of the methodology and data used, and caution in interpreting results when carrying out program evaluations.

Project Description. PROBECAT (Programa de Becas de Capacitacion para Trabajadores) is a Mexican short-term training program targeted at increasing earnings and employment for unemployed and displaced workers. PROBECAT is administered through the state employment offices. Trainees receive minimum wage during the training period, which lasts from one to six months, and the local employment office provides placement. Originally, the program was small (50,000 or so participants), but in recent years it has grown dramatically, to cover more than 500,000 persons per year.

Highlights of the Evaluations. The highlights are as follows:

- The 1994 evaluation is interesting for four reasons: the imaginative use of existing data; the construction of a matched-comparison group; the explicit recognition of the multifaceted nature of the intervention outcomes, particularly for heterogeneous groups of workers; and the explicit cost-benefit analysis. The findings of the evaluation were quite positive in terms of the impact of the program on beneficiaries.
- The 1995 evaluation is a replication of the methodology of the 1994 evaluation on a more recent data set. The findings are also favorable for the impact of the program. Because the design and findings of the 1995 evaluation match those of the 1994 evaluation, the 1995 evaluation will not be discussed below.

- The 1999 evaluation was carried out as part of the Mexico poverty assessment with the data set used for the 1995 evaluation but with a different econometric methodology. The controls used for the endogeneity of program participation showed a vanishing of the impact of the program on the probability of working and on wages after training. Although this does not imply that the program has no benefit, it suggests that it works more as a temporary safety net for the unemployed than as a job training program.

II. Research Questions and Evaluation Design

In the 1994 evaluation, the authors estimate the impact of training on (a) the probability of employment after 3, 6, and 12 months; (b) the time to exit unemployment; (c) the effect on monthly earnings, work hours per week, and hourly wages; and (d) the return on investment.

The 1999 evaluation looks at the same questions except work hours per week and hourly wages. Given that there is no impact in that evaluation on employment and monthly earnings, the return is zero, but again the program may work as a safety net.

The design of both evaluations is innovative in constructing the comparison group. In both cases, the evaluations combine an existing panel labor force survey, Encuesta Nacional de Empleo (ENEU), with a panel of trainees for the same period. That is, the program's selection criteria are used to define the control group from the ENEU. Although there is no alternative to this combination of surveys because of data limitations, the construction of the joint sample (control and treatment groups) can be critiqued, as discussed in the 1999 evaluation:

- In using the unemployed individuals in the ENEU to form the control group, it is assumed that none of the ENEU individuals have benefited from the program. This is not the case because every individual in the ENEU has some probability of having participated in PROBECAT. Fortunately, given that the program was small until 1993, only a very small minority of the individuals in the control group are likely to have participated in the program (the data for the 1999 evaluation are for 1993–94);
- The combination of two random samples (PROBECAT trainees and ENEU unemployed individuals) is not a random sample, so that in the absence of the standard properties for the residuals, the results of regressions may not yield consistent parameter estimates, especially because the models used are sensitive to the assumption of bivariate normality. In the absence of better data, not much can be done on this.

The main differences between the 1994 and 1999 evaluations are as follows;

- In the 1994 evaluation, the authors attempt to address the selection bias problems resulting from PROBECAT's nonrandom selection of trainees by estimating a probit model of the probability of participation. The comparison group is then limited to those individuals who are highly likely to participate. In the 1999 evaluation, the authors argue that this method does not eliminate the problem of endogeneity. Instead, they use an instrumental variable to control for the endogeneity of program participation.
- In the estimation of earnings in the 1994 evaluation, while participation in PROBECAT is controlled for, the sample selection bias resulting from the decision to work is not accounted for. In the 1999 study, both sample selection problems are accounted for.

III. Data

In the 1994 evaluation, data on trainees are gathered from a 1992 retrospective survey administered to 881 men and 845 women who were trained in 1990. This is supplemented with panel data on 371 men and 189 women derived from a household survey of the 16 main urban areas in Mexico. This survey was part of a regular quarterly labor force survey, ENEU, undertaken by the Mexican statistical agency. The authors exploited the rotation group structure of the survey to take workers who were unemployed in the third quarter of 1990 and then tracked those workers for a year. This was supplemented by a cohort that became unemployed in the fourth quarter of 1990 and was tracked for nine months. The same method was used in the 1999 evaluation, but for more recent data.

IV. Econometric Techniques

The key econometric techniques used are survival analysis (duration models) for the probability of working and Heckman regressions for wages. What follows is based on the 1999 evaluation. Differences with the 1994 evaluation are highlighted.

Impact of PROBECAT on the Length of Employment Search. In the survival analysis, the survivor function $S(t)$ represents the length of unemployment after training (measured in months). Given $S(t)$, the hazard function $\lambda(t)$ denoting the chance of becoming employed (or the risk of remaining unemployed) at time t among the individuals who are not yet employed at that time is $\lambda(t) = -d(\log S(t))/dt$. The survivor curve can be

specified as a function of program participation P , individual characteristics X , and state characteristics Z , so that $\lambda = \lambda(t; X, Z, P)$. In Cox's proportional hazard model, if i denotes a household and j denotes the area in which the household lives, we have

$$\lambda(t; X, Z, P1, P2) = \lambda_0(t) \exp(\gamma'X_{ij} + \delta'Z_j + \mu P_{ij}). \quad (1)$$

Cox proposed a partial maximum likelihood estimation of this model in which the baseline function $\lambda_0(t)$ does not need to be specified. If μ is positive and statistically significant, the program has a positive effect on employment. In a stylized way, the difference between the 1994 and 1996 evaluations can be described as follows:

- In the 1994 evaluation, the authors run a probit on program participation and delete from the control group those individuals with a low probability of participating in the program. They then run equation (1) without further control for endogeneity.
- In the 1999 evaluation, the authors also run a probit on program participation, but they use program availability at the local level (obtained from administrative data) as an additional determinant of participation (but not of outcome conditional on individual participation.) Then they run equation (1), not with the actual value of the participation variable but with the predicted (index) value obtained from the first stage probit. This is an instrumental variable procedure. The idea follows work on program evaluation using decentralization properties by Ravallion and Wodon (2000) and Cord and Wodon (1999). The authors compare their results with other methods, showing that other methods exhibit a bias in the value of the parameter estimates owing to insufficient control for endogeneity.

Impact of PROBECAT on Monthly Earnings. To carry out this analysis, a model with controls for sample selection in labor force and program participation is used in the 1999 evaluation (the 1994 evaluation controls only for program participation). Denote by $\log w$ the logarithm of the expected wage for an individual. This wage is nonzero if and only if it is larger than the individual's reservation wage (otherwise, the individual would choose not to work). Denote the unobserved difference between the individual's expected wage and his or her reservation wage by Δ^* . The individual's expected wage is determined by a number of individual (vector E , consisting essentially of the individual's education and past experience) and geographic variables Z , plus program participation P . The difference between the individual's expected wage and his or her reservation wage is determined by the same variables, plus the number

of children, the fact of being a household head, and the fact of being married, captured by D . The model is thus

$$\Delta_{ij}^* = \phi_{\Delta}'E_{ij} + \pi_{\Delta}'D_{ij} + \eta_{\Delta}'Z_j + \alpha_{\Delta}P_{ij} + v_{ij} \text{ with } \Delta_{ij} = 1 \text{ if } \Delta_{ij}^* > 0, \text{ and } 0 \text{ if } \Delta_{ij}^* < 0 \quad (2)$$

$$\text{Log } w_{ij}^* = \phi_w'E_{ij} + \eta_w'Z_j + \alpha_w P + \kappa_{ij} \text{ with } \text{Log } w = \text{log } w^* \text{ if } \Delta = 1 \text{ and } 0 \text{ if } \Delta = 0. \quad (3)$$

As for the survival model, in order to control for endogeneity of program participation, in the 1999 evaluation a probit for program participation is first estimated by using program availability at the local level as a determinant of individual participation. Then the above equations are estimated by using the predicted (index) value of program participation instead of its true value. In the 1994 evaluation, the model does not control for the decision to participate in the labor market given in equation (2) above. This equation is replaced by the program participation probit estimated without local availability of the program as an independent variable. Again, comparisons of various models show that bias is present when the instrumental variable technique is not used.

V. Who Carried It Out

The 1994 evaluation was conducted by Ana Revenga in the Latin America and Caribbean Country Department II of the World Bank, Michelle Riboud in the Europe and Central Asia Country Department IV of the World Bank, and Hong Tan in the Private Sector Development Department of the World Bank. The 1999 evaluation was carried out by Quentin Wodon and Mari Minowa, also at the World Bank (Latin America region).

VI. Results

The results obtained in the various evaluations are very different. The 1994 and 1995 evaluations find positive impacts of the program on employment and wages. No positive impact was found in the 1999 evaluation, which is based on the same data used for the 1995 evaluation. In terms of cost-benefit analysis, the first two evaluations are favorable but the last evaluation is not. The disappointing results in the last evaluation are not surprising. Most retraining programs in Organisation for Economic Co-operation and Development countries have been found to have limited impacts, and when programs have been found to have some impact, this impact tends to vanish after a few years (Dar and Gill 1998).

The fact that PROBECAT may not be beneficial in the medium to long run for participants according to the last evaluation does not mean that it should be suppressed. The program could be viewed as providing temporary safety nets (through the minimum wage stipend) rather than training. Or it could be improved so as to provide training with longer-lasting effects.

VII. Lessons Learned

Apart from some of the innovative features of these evaluations and their limits, the key lesson is that one should be very careful in doing program evaluations and using the results to recommend policy options. The fact that a subsequent evaluation may contradict a previous one with the use of different econometric techniques should always be kept in mind. There have been many such cases in the literature.

VIII. Sources

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

Wodon, Quentin, and Mari Minowa. "Training for the Urban Unemployed: A Reevaluation of Mexico's PROBECAT." World Bank, Government Programs and Poverty in Mexico, Report No. 19214-ME, Vol. II.

Annex 1.10: Mexico, National Program for Education, Health, and Nutrition (PROGRESA): A Proposal for Evaluation

I. Introduction

Project Description. PROGRESA is a multisectoral program aimed at fighting extreme poverty in Mexico by providing an integrated package of health, nutrition, and educational services to poor families. The Mexican government will provide monetary assistance, nutritional supplements, educational grants, and a basic health package for at least three consecutive years. It plans to expand PROGRESA from its current size of 400,000 families to 1 to 1.5 million families at the end of 1998, with an expenditure of \$500 million.

Highlights of Evaluation. The evaluation is particularly complex because three dimensions of the program are evaluated: operation, targeting effectiveness, and impact. Adding to the complexity, outcomes are themselves multidimensional. There are thus many different evaluation components: beneficiary selection, evaluation methods, nonexperimental analytical framework, data requirements, impacts on education, impacts on health, impacts on food consumption and nutrition, impacts on consumption expenditures and intrahousehold allocation, potential second-round impacts of the program, simulations of changes in program benefits, and cost-effectiveness and cost-benefit issues.

Although the evaluation is an outline of ideas rather than the results of an implementation, a major lesson learned from it is how to think about and structure an evaluation before actually implementing it. In particular, there is a very useful outline of the conceptual and empirical issues to be addressed in an evaluation and the ways in which the issues can be addressed. Another useful component of the evaluation is its breadth: rather than simply evaluating the impact of an intervention, it will help pinpoint whether the outcome is due to successes or failures in the intervention operation and targeting.

II. Research Questions and Evaluation Design

The core research questions are to evaluate the three dimensions of PROGRESA's performance—operational aspects, targeting, and impact. The operational aspect of an intervention is often ignored, despite the fact that interventions could be turned from failures into successes if correc-

tive measures were taken. A similar argument could be made for targeting: a program may seem to have failed simply because of poor targeting rather than because the intervention itself was flawed. The evaluation of the impact is more standard, although even this goal is quite ambitious in that both the magnitude of the impact and the pathways by which it is achieved are analyzed.

The monitoring of the program operation is a two-step procedure. The team develops a schematic of the sequence of steps for the intervention. The team then uses observations, interviews, focus groups, and workshops with stakeholders to assess, analyze, and potentially change program processes.

A two-step approach is also used to target households for PROGRESA. The first is to identify which localities in a region are eligible to receive PROGRESA by means of a poverty-based index. The second is to identify the eligibility of a family within the locality, based on the interaction between PROGRESA officials and local leaders. The study will address the validity of this targeting by (a) comparing the distribution of household consumption levels in participant and nonparticipant households in treatment localities, (b) deriving an eligibility cutoff for household consumption that is consistent with the total number of households that PROGRESA can serve, (c) conducting sensitivity and specificity analysis of PROGRESA and non-PROGRESA households versus the households selected and not selected under this cutoff, (d) exploring the ability of current criteria to predict consumption, (e) identifying alternative criteria from other data sources, and (f) simulating models that could improve targeting with alternative criteria (International Food Policy Research Institute 1998, p. 6).

For the impact evaluation, the same system was followed, with the result that localities were randomly allocated to 296 treatment and 173 nontreatment groups, with 14,382 families in the former category and 9,202 families in the latter category. Eligible families in the control category will receive treatment after at least one year has passed.

The consultants plan to test for possible nonrandomization by comparing the characteristics of treatment and control groups. If they are systematically different, then three nonexperimental methods will be used: control function methods, matching methods, and regression methods.

III. Data

The operational data component is obtained from observation and interviews, focus groups, and workshops with stakeholders. The main focus is on identifying what and why things are happening, the level of satis-

faction with the process, and improvement suggestions. These data are collected across localities and will also rely heavily on PROGRESA's internal administrative records.

Two surveys have been implemented: December 1997 census surveys and March 1998 baseline surveys. The central variable for the targeting criterion is clearly household consumption, and while this was not collected in the census, it was collected in the March survey. This variable, however, lacks information on self-consumption, and although it will be collected later, it will be contaminated by the implementation of PROGRESA. The consultants plan to work exclusively with eligible and noneligible households in the control localities.

The evaluation of the impact hinges on the choice of impact indicators. PROGRESA should affect both the quality and quantity of services provided and investment in health, nutrition, and education. A host of evaluation indicators are proposed based on a number of impact outcomes, and each has an associated data source. Household welfare, as measured by household consumption, savings, accumulation of durable goods, will be measured by baseline and follow-up surveys; the nutritional and health status of children will be measured by a nutrition subsample baseline and follow-up surveys; child educational achievement will be measured by standardized national tests; food consumption will be captured by the baseline and follow-up surveys; school use will be addressed by both a school-level survey and by the baseline and follow-up surveys; health facility use can be monitored by health clinic records and the surveys; and women's status can also be measured by surveys and by the stakeholder investigations.

One very attractive feature of the proposed evaluation is the analytical approach taken to examine current outcome measures and the extensive discussion of more appropriate outcome and control measures for education, health, and consumption.

A cost-benefit analysis is planned. A set of benefits is developed, despite the inherent difficulty of monetizing quality of life and empowerment improvements. Two different types of cost are also identified: administrative program costs and program costs. The former consist of screening, targeted delivery mechanisms, and monitoring costs; the latter include forgone income generation.

IV. Econometric Techniques

The econometric techniques applied depend on the relationships to be estimated. The consultants discuss the appropriateness of the production function relationship (for example, for academic achievement), demand relationships (for example, for health or education services), and condi-

tional demand relationships (in which some variables are determined by the family rather than the individual).

The most interesting econometric technique used is applied to the estimation of a Working-Leser expenditure function of the form

$$W_j = \alpha_1 + \beta_{1j} \text{lpexp} + \beta_{2j} \text{lsiz} + \sum_k \delta_{kj} \text{dem}_k + \sum_s \Theta_{sj} z_s + \beta_{3j} P + e_j$$

where w_j is the budget share of the j th good; lpexp is the log of per capita total expenditures; lsiz is the log of household size; dem_k is the proportion of demographic group k in the household; z_s is a vector of dummy variables affecting household location; P captures Progresa participation; and e_j is the error term.

This approach has many advantages: it permits the inclusion of control factors; it satisfies the adding-up constraint; and it is widely used, permitting comparisons with other studies. Finally, the model can be used to identify three different paths in which PROGRESA can affect expenditures: through changing household resources (β_{1j} times the marginal propensity to consume, estimated separately), through changing the income distribution (by modifying it to include the proportion of adult women in the household), and through a greater participation effect. The baseline and follow-up surveys allow difference-in-difference methodologies to be used.

They also identify key econometric issues that are likely to be faced: collinearity, measurement error, omitted variables, simultaneity, and identifying the time period within which it is reasonable to expect an impact to be observable.

V. Who Will Carry It Out

The International Food Policy Research Institute staff include Gaurav Datt, Lawrence Haddad, John Hoddinott, Agnes Quisumbing, and Marie Ruel. The team includes Jere Behrman, Paul Gertler, and Paul Schultz.

VI. Lessons Learned

The primary lesson learned here is the value of identifying evaluation issues, methodology, and data sources—and critically evaluating the evaluation—before the evaluation takes place. This evaluation outline provides a very valuable service in developing a thoughtful illustration of all the possible issues and pitfalls an evaluator is likely to encounter. In particular, some common-sense issues with evaluating an impact are identified: (a) policy changes may be hard to predict because of cross-substitution and behavior adjustment; (b) marginal benefits and margin-

al costs depend on a number of things: externalities (putting a wedge between social and private valuation), the actors (parents versus children); (c) the importance of unobserved characteristics; (d) the importance of controlling for individual, family, and community characteristics; and (e) the empirical estimates depend on a given macroeconomic, market, policy, and regulatory environment.

VII. Source

International Food Policy Research Institute. 1998. *Programa Nacional de Educación, Salud, y Alimentación (PROGRESA): A Proposal for Evaluation* (with technical appendix). Washington, D.C.: IFPRI.

Annex 1.11: Evaluating Nicaragua's School Reform: A Combined Quantitative-Qualitative Approach

I. Introduction

Project Description. In 1991, the Nicaraguan government introduced a sweeping reform of its public education system. The reform process has decentralized school management (decisions on personnel, budgets, curriculum, and pedagogy) and transferred financing responsibilities to the local level.

Reforms have been phased in over time, beginning with a 1991 decree that established community-parent councils in all public schools. Then a 1993 pilot program in 20 hand-picked secondary schools transformed these councils into school management boards with greater responsibility for personnel, budgets, curriculum, and pedagogy. By 1995, school management boards were operational in 100 secondary schools and over 300 primary schools, which entered the program through a self-selection process involving a petition from teachers and school directors. School autonomy was expected to be almost universal by the end of 1999.

The goal of the Nicaraguan reforms is to enhance student learning by altering organizational processes within public schools so that decision-making benefits students as a first priority. As school management becomes more democratic and participatory and locally generated revenues increase, spending patterns are to become more rational and allocated to efforts that directly improve pedagogy and boost student achievement.

Impact Evaluation. The evaluation of the Nicaraguan School Autonomy Reform represents one of the first systematic efforts to evaluate the impact of school decentralization on student outcomes. The evaluation, carried out jointly by the World Bank and the Ministry of Education, began in 1995 and was to be complete by the end of 1999. The design is innovative in that it combines both qualitative and quantitative assessment methods, and the quantitative component is unique in that it includes a separate module assessing school decisionmaking processes. The evaluation also illustrates "best-practice" techniques when there is no baseline data and when selective (nonrandom) application of reforms rules out an experimental evaluation design.

The purpose of the qualitative component of the evaluation is to illuminate whether or not the intended management and financing reforms

are actually observed in schools and to assess how various stakeholders viewed the reform process. The quantitative component fleshes out these results by answering the following question: "Do changes in school management and financing actually produce better learning outcomes for children?" The qualitative results show that successful implementation of the reforms depends largely on school context and environment (i.e., poverty level of the community), whereas the quantitative results suggest that increased decisionmaking by schools is in fact significantly associated with improved student performance.

II. Evaluation Design

The design of the Nicaraguan School Autonomy Reform evaluation is based on the "matched comparison technique," in which data for a representative sample of schools participating in the reform process are compared with data from a sample of nonparticipating schools. The sample of nonparticipating schools is chosen to match, as closely as possible, the characteristics of the participating schools and hence provides the counterfactual. This design was chosen because the lack of baseline data ruled out a before-and-after evaluation technique and because reforms were not applied randomly to schools, which ruled out an experimental evaluation design (in which the sample of schools studied in the evaluation would be random and therefore nationally representative).

III. Data Collection and Analysis Techniques

The qualitative study draws on data for a sample of 12 schools, 9 reformers and 3 nonreformers, which represent the control group. (Data were actually gathered for 18 schools, but only 12 of these schools were included in the qualitative study because of delays in getting the transcripts prepared and a decision to concentrate the bulk of the analysis on reform schools, which provided more relevant material for the analysis.) The sample of 12 schools was picked to represent both primary and secondary schools, rural and urban schools, and, based on data from the 1995 quantitative survey, schools with differing degrees of actual autonomy in decisionmaking. A total of 82 interview and focus-group sessions were conducted, focusing on discovering how school directors, council members, parents, and teachers understood and viewed the decentralization process. All interviews were conducted by native Nicaraguans, trained through interview simulation and pilot tests to use a series of guided questions without cueing responses. Interviews were audio-recorded, transcribed, and then distilled into a two- to four-page transcript, which was then analyzed to identify discrete sets of evidence and

fundamental themes that emerged across schools and actors and between reform schools and the control group.

Quantitative data collection consisted of two components, a panel survey of schools that was conducted in two rounds (November–December 1995 and April–August 1997) and student achievement tests for students in these schools that were conducted in November 1996. The school survey collected data on school enrollment, repetition and dropout rates, physical and human resources, school decisionmaking, and characteristics of school directors, teachers, students, and their families. The school decisionmaking module is unique and presents a series of 25 questions designed to gauge whether and how the reform has actually increased decisionmaking by schools. The survey covered 116 secondary schools (73 reformers and 43 nonreformers representing the control group) and 126 primary schools (80 reformers and 46 nonreformers). Again, the control groups were selected to match the characteristics of the reform schools. The survey also gathered data for 400 teachers, 182 council members, and 3,000 students and their parents, and 10–15 students were chosen at random from each school. Those students who remained in school and could be traced were given achievement tests at the end of the 1996 school year and again in the second round of survey data collection in 1997.

Quantitative data analysis draws on regression techniques to estimate an education production function. This technique examines the impact of the school's management regime (how decentralized it is) on student achievement levels, controlling for school inputs, and household and student characteristics. The analysis measures the effect of both *de jure* and *de facto* decentralization; *de jure* decentralization simply indicates whether or not the school has legally joined the reform, whereas *de facto* decentralization measures the degree of actual autonomy achieved by the school. *De facto* decentralization is measured as the percentage of 25 key decisions made by the school itself and is expected to vary across schools because reforms were phased in (so schools in the sample will be at different stages in the reform process) and because the capacity to successfully implement reforms varies according to school context (a result identified in the qualitative study).

IV. Results

The qualitative study points out that policy changes at the central level do not always result in tidy causal flows to the local level. In general, reforms are associated with increased parental participation as well as management and leadership improvements. But the degree of success with which reforms are implemented varies with school context. Of par-

ticular importance are the degree of impoverishment of the surrounding community (in poor communities, increasing local school financing is difficult) and the degree of cohesion among school staff (when key actors such as teachers do not feel integrated into the reform process, success at decentralization has been limited). Policymakers often ignore the highly variable local contexts into which new programs are introduced. The qualitative results point out that in the Nicaraguan context the goal of increased local financing for schools is likely to be derailed in practice, particularly in poor communities, and therefore merits rethinking.

The quantitative study reinforces the finding that reform schools are indeed making more of their own decisions, particularly with regard to pedagogical and personnel matters. *De jure* autonomy—whether a school has signed the reform contract—does not necessarily translate into greater school-level decisionmaking, or affect schools equally. The degree of autonomy achieved depends on the poverty level of the community and how long the school has been participating in the reform process. The regression results show that *de jure* autonomy has little bearing on student achievement outcomes; but *de facto* autonomy—the degree of actual decentralization achieved by the school—is significantly associated with improved student achievement. (This result is preliminary pending further exploration of the panel data, which have recently become available.) Furthermore, simulations indicate that increased school decentralization has a stronger bearing on student achievement than improvements in other indicators of typical policy focus, such as teacher training, lowering class size, and increasing the number of textbooks.

V. Policy Application

The evaluation results provide concrete evidence that Nicaragua's School Autonomy Reform has produced tangible results. Reform schools are indeed making more decisions locally—decentralization is happening in practice, not just on the books—and enhanced local decisionmaking does result in improved student achievement.

The results also point out areas in which policy can be improved, and, as a result, the Ministry of Education has introduced a number of changes in the school reform program. The program now places greater emphasis on the role of teachers and on promoting the pedagogical aspects of the reform. Teacher training is now included as part of the program, and the establishment of a pedagogical council is being considered. Further, in response to the financing problems of poor communities, the ministry has developed a poverty map-driven subsidy scheme. Finally, the tangible benefits from this evaluation have prompted the ministry to incorporate a permanent evaluation component into the reform program.

VI. Evaluation Costs and Administration

Costs. The total cost of the evaluation was approximately \$495,000, representing less than 1.5 percent of the World Bank credit. (This total does not include the cost of local counterpart teams in the Nicaraguan Ministry of Education.) Of this total evaluation cost, 39 percent was spent on technical support provided by outside consultants, 35 percent on data collection, 18 percent on World Bank staff time, and 8 percent on travel.

Administration. The evaluation was carried out jointly by the Nicaraguan Ministry of Education and the World Bank. In Nicaragua the evaluation team was led by Patricia Callejas, Nora Gordon, and Nora Mayorga de Caldera in the Ministry of Education. At the World Bank the evaluation was carried out as part of the research project, "Impact Evaluation of Education Projects Involving Decentralization and Privatization" under the guidance of Elizabeth King, with Laura Rawlings and Berk Ozler. Coordinated by the World Bank team, Bruce Fuller and Madgalena Rivarola from the Harvard School of Education worked with Liliam Lopez from the Nicaraguan Ministry of Education to conduct the qualitative evaluation.

VII. Lessons Learned

Value of the Mixed-Method Approach. Using both qualitative and quantitative research techniques generated a valuable combination of useful, policy relevant results. The quantitative work provided a broad, statistically valid overview of school conditions and outcomes; the qualitative work enhanced these results with insight into why some expected outcomes of the reform program had been successful whereas others had failed and hence helped guide policy adjustments. Furthermore, because it is more intuitive, the qualitative work was more accessible and therefore interesting to ministry staff, which in turn facilitated rapid capacity building and credibility for the evaluation process within the ministry.

Importance of Local Capacity Building. Local capacity building was costly and required frequent contact and coordination with World Bank counterparts and outside consultants. However, the benefit was the rapid development of local ownership and responsibility for the evaluation process, which in turn fostered a high degree of acceptance of the evaluation results, whether or not these reflected positively or negatively on the program. These evaluation results provided direct input to the reform as it was evolving. The policy impact of the evaluation was also enhanced by a cohesive local team in which evaluators and policymakers worked

collaboratively, and because the minister of education was brought on board as an integral supporter of the evaluation process.

VIII. Sources

The following documents provide detailed information on the Nicaraguan School Autonomy Reform Evaluation:

Fuller, Bruce, and Magdalena Rivarola. 1998. *Nicaragua's Experiment to Decentralize Schools: Views of Parents, Teachers and Directors*. Working Paper Series on Impact Evaluation of Education Reforms, paper no. 5. World Bank, Washington, D.C.

King, Elizabeth, and Berk Ozler. 1998. *What's Decentralization Got to Do with Learning? The Case of Nicaragua's School Autonomy Reform*. Working Paper Series on Impact Evaluation of Education Reforms, paper no. 9. World Bank, Washington, D.C.

King, Elizabeth, Berk Ozler, and Laura Rawlings. 1999. *Nicaragua's School Autonomy Reform: Fact or Fiction?* Washington, D.C.: World Bank.

Nicaragua Reform Evaluation Team. 1996. *Nicaragua's School Autonomy Reform: A First Look*. Working Paper Series on Impact Evaluation of Education Reforms, paper no. 1. World Bank, Washington, D.C.

Nicaragua Reform Evaluation Team. 1996. *1995 and 1997 Questionnaires, Nicaragua School Autonomy Reform*. Working Paper Series on Impact Evaluation of Education Reforms, paper no. 7. World Bank, Washington, D.C.

Rawlings, Laura. 2000. "Assessing Educational Management and Quality in Nicaragua." In Bamberger, *Integrating Quantitative and Qualitative Methods in Development Research*. Washington, D.C.: World Bank.

Annex 1.12: Improving Elementary Mathematics Education in Nicaragua: An Experimental Study of the Impact of Textbooks and Radio on Achievement

I. Summary of Evaluation

Most poor countries have extremely limited resources for education, which makes it important to allocate those resources effectively. Of the three common policy options available—smaller class sizes, longer teacher training programs, and textbook provision—only the last has frequently been found to have a significantly positive effect on student learning. This evaluation quantified the impact of textbook availability on mathematics learning for Nicaraguan first grade students.

The design of the evaluation was to provide textbooks to all students in a subset of classes that were originally designated to be controls in an ongoing study of the effectiveness of radio instructional programs. Half of the classes received textbooks; half did not. All classes received both a pretest at the beginning of the year and a posttest at the end. The study then used simple regression techniques to compare the mean classroom posttest scores as a function of pretest scores and the intervention.

A major lesson learned is how to carefully design an evaluation: the randomization was particularly well-constructed and cleverly combined with a test that maximized cross-class comparability. Another lesson learned was one of pragmatism: the evaluation was designed to forestall potentially quite serious political economy issues. Finally, the evaluation provides a series of practical examples of the types of decisions that must be made in fieldwork.

II. Research Questions and Evaluation Design

There are two very interesting components of the evaluation design: the piggy-backing on a preexisting evaluation and the up-front understanding of the political environment within which the evaluation was to take place. The key research question was straightforward: to assess the impact of increased textbook availability on first grade student learning—particularly focusing on whether the textbooks were actually used in the classroom. Because there was already a radio instructional program intervention (Radio Mathematics) in place, the question was broadened to compare the impact of textbook availability with radio instruction as well as with a control group.

It is worth discussing the decision to monitor the actual use of textbooks, which makes the evaluation more difficult. Many educational interventions provide materials to classrooms, but clearly the impact of the provision depends on use. However, as the evaluators point out, this decision means that the evaluation “does not assess the potential that textbooks or radio lessons have for improving student achievement under optimal outcomes. Rather, it attempts to assess their impact as they might be adopted in the typical developing country” (Jamison, 1981 p. 559). Thus simple textbook provision may not in itself suffice without also designing a method to ensure that teachers use the textbooks as intended.

The evaluation used a randomized design that was piggybacked on a preexisting project evaluation. In the existing Radio Nicaragua Project, an entire mechanism had already put random assignment and testing procedures in place in order to evaluate the effectiveness of a radio-based instructional program. The existing project had already classified all primary schools in three provinces in Nicaragua as radio or control using a random sampling process stratified by urbanization (about 30 percent of students are in rural schools, but equal numbers of classes were chosen in each stratum).

The textbook evaluation exploited this preexisting design by selecting treatment and control schools in the following fashion. First, the evaluators acquired a list of all schools with eligible classrooms for each of the six categories (three provinces, rural and urban). They then randomly assigned schools to treatment or control from these master lists for each category, and then schools were used in the order that they appeared (one school, which refused to participate, was replaced by the next one on the list). Requests to participate from classes in control groups were denied, and all use of the experimental material was controlled by the authors. It is useful to note that the evaluation design had addressed this potential political difficulty up front. The evaluation team announced their intentions from the outset; the team obtained official approval and support of the policy, and the team also established clear and consistent procedures for the program.

The study thus randomly selected 88 classrooms: 48 radio and 40 control schools. Twenty of the control schools received textbooks for each child, and teachers received both written and oral instruction and the teachers' editions of the tests. The radio component consisted of 150 daily mathematics lessons, combined with student worksheets and written and oral teacher instructions.

An interesting decision that was made was the deliberate lack of supervision of treatment groups. This was clearly difficult because the absence of supervision made it hard to assess program utilization.

However, the cost in terms of influencing behavior was judged to be too high. Surprise visits, which were the accepted compromise solution, could not be used because of political turmoil during the assessment year and so had to be conducted the following year.

A second decision was to have tests administered by project staff rather than classroom teachers. This clearly increased administrative costs but reduced potential bias in test taking. The students were given a pretest of mathematical readiness during the first three weeks of school. The posttest, which measured achievement, was intended to be given in the last three weeks of school but was administered two weeks early because of political problems. The students had, as much as possible, identical conditions for both tests when they took them because they had the same length of time for the tests and because instructions were taped.

III. Data

There are two main lessons to be drawn from the data collection component. The first is that logistical difficulties are often inevitable. Despite the careful design there were a series of problems with developing a perfect set of pretest-posttest comparisons. Although there were a total of 20 control classes, 20 textbook classes, and 47 radio classes, the numbers of pretest and posttest scores were different in each group because of late registration, dropping out, absence, and failure to be tested because of overcrowding. Individual information on the students does not appear to have been collected.

The second lesson is the imaginative way in which the evaluators designed the posttest to minimize burden and yet obtain the necessary information. A series of issues were faced:

- There were no standardized tests in use in Nicaragua.
- The test had to assess the achievement of the curriculum objectives.
- The test had to capture achievement on each topic to facilitate an evaluation of the effectiveness of the intervention on each topic as well as in total.

The evaluators used a multiple matrix-sampling design to address these issues. The test had two types of questions: those given to all the students in the class (40 G items) and those given to subsets of students (44 I items). All I items were tested in every classroom; one-quarter of all G items were tested in each classroom. This enables the researchers to randomly assign units across two dimensions: schools and test forms. The mean posttest scores for treatment and control groups are derived by adding average scores for each test, and the standard errors are calculat-

ed by using the residual variance after removing the main effects of items and students.

Information on textbook usage was also collected the year after the intervention from 19 of the 20 textbook-using schools.

IV. Econometric Techniques

The structure of the evaluation meant that a simple comparison of means between treatment and control groups would be appropriate, and this was in fact used. The approach can be very cumbersome if there are multiple strata and multiple interventions, which was the case with this evaluation. Thus the evaluators also used a simple regression approach. Here the class was the unit of analysis, and the class mean posttest score was regressed against the mean pretest score as well as dummies for the radio and textbook interventions, an urban-rural dummy, and the average class pretest score as independent variables.

An important component of any evaluation is whether different groups are affected differently by the same treatment. This can often be achieved, as was done in this evaluation, by imaginative use of interactive variables. Differences between urban and rural areas were captured by interacting the urban-rural dummy with the intervention; difference in the effect of the intervention based on initial test scores was captured by interacting initial test scores with the intervention.

V. Who Carried It Out

The World Bank supported the research project, but it was imbedded in the joint United States Agency for International Development–Nicaragua Ministry of Education Radio Mathematics Project.

VI. Results

The authors found that both textbook and radio treatments had important effects on student outcomes: textbook availability increased student posttest scores by 3.5 items correct, radio lessons by 14.9 items—quite substantial given that the classroom standard deviation is 8.3 and that of individual items is 11.8. Radio lessons and textbooks were both more effective in rural schools and could potentially play a large part in reducing the gap between urban and rural quality. These results appear to be independent of the initial skill level of the class, as measured by pretest scores.

The authors attribute the difference in outcomes for the radio and the textbook interventions to differences in textbook usage, particularly given poorly educated teachers.

VII. Lessons Learned

Three main lessons were learned: the importance of politics in design decisions, the usefulness of imaginative test designs, and the difficulties associated with fieldwork. First, the political economy of randomized design was highlighted in this study: there are clearly quite strong political pressures that can be brought to bear and that need to be addressed early on and with the support of the government. Second, the authors were able to measure many facets of learning outcomes without having unrealistically long tests, by imaginative application of a test design. Finally, the evaluators clearly addressed a number of fieldwork questions: whether and how to monitor the actual adoption of textbooks and who should administer the tests.

VIII. Source

Jamison, Dean T., Barbara Serle, 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.

Annex 1.13: The Impact of Alternative Cost-Recovery Schemes on Access and Equity in Niger

I. Introduction

Project Description. The ability to recover some portion of health care costs is critical to the provision of health care. Little is known, however, about the effect of different strategies on quality and welfare outcomes. The evaluation estimates the impact on the demand for health care of two pilot cost-recovery schemes in the primary care (nonhospital) sector in Niger. Niger is a poor, rural economy; public health costs are 5 to 6 percent of the government budget; and much of this financing is mistargeted toward hospitals and personnel. The government wanted to evaluate the consequences of different payment mechanisms and considered two: a pure fee-for-service and a tax plus fee-for-service financing mechanism, both of which were combined with quality and management improvements. The government was particularly interested in finding out how the demand for health care changed, particularly among vulnerable groups, and in examining whether such quality improvements were sustainable.

Highlights of Evaluation. The different payment mechanisms were implemented in three districts, one for each treatment and one control. The evaluation used a quasi-experimental design based on household surveys combined with administrative data on utilization and operating costs. The evaluation is particularly attractive in that it directly addresses political economy issues with a survey instrument that asks respondents about their willingness to pay for the improved service. This explicit recognition that significant outcomes are not, by themselves, enough to guarantee a sustainable project is an extremely valuable contribution. Another useful aspect is the explicit evaluation of the impact of the intervention on different target groups (children, women, villages without a public health facility, and the poorest citizens).

II. Research Questions and Evaluation Design

The main questions were the impact of the treatment on (a) the demand for and utilization of public health care facilities, (b) specific target groups (poor, women, and children), (c) financial and geographic access, (d) the use of alternative services, and (e) the sustainability of improvements under cost recovery (patient and drug costs as well as revenues and willingness to pay).

Three health districts were selected in different provinces from an administrative register. Although all were similar in terms of economic, demographic, and social characteristics, they were ethnically different. Each district had a medical center, with a maternal and child health center, one medical post, and one physician as well as rural dispensaries.

Four quality and management improvements were instituted in the two treatment districts; none was implemented in the control district. In particular, initial stocks of drugs were delivered; personnel were trained in diagnosis and treatment; a drug stock and financial management system was installed and staff were trained in its use; supervisory capacity was increased to reinforce management.

The two different pricing mechanisms were introduced at the same time. The first was a fee-per-episode, with a fee of 200 FCFA (US\$0.66) for a user over age five, a fee of 100 FCFA (US\$0.33) for a user under five. The second combined an annual tax of 200 FCFA paid by district taxpayers and a fee of 50 FCFA per user over five and 25 FCFA for children under five. Annual income was under US\$300 per capita. Each scheme included exemptions for targeted groups. The funds were managed at the district level.

III. Data

The three districts were chosen from administrative data. Two household surveys were implemented, one of which was a baseline, and these were combined with administrative records on facilities. Each survey collected demographic household and individual information from a randomly selected sample of 1,800 households. The baseline survey had information on 2,833 individuals who had been sick the two weeks before the survey and 1,770 childbearing women; the final survey had data on 2,710 sick individuals and 1,615 childbearing women. The administrative data consisted of quite detailed information on monthly expenditures on drug consumption and administration, personnel maintenance, and fee receipts together with the utilization of the health facilities. This information was collected in the year before the intervention, the base year (May 1992–April 1993), and the year after the intervention.

IV. Econometric Techniques

The study combines comparisons of means with simple logit techniques, the latter being used to capture utilization changes. In particular, the individual response of whether the health care facility was used (PI) to specify the following model:

$$\text{Logit}(P_i) = X\beta + \alpha(A + B).$$

This model, which controls for a vector of individual characteristics X as well as dummy variables A and B , was compared with

$$\text{Logit}(P_i) = X\beta + \alpha_a A + \alpha_b B.$$

The dummy variables A and B are variously defined. In the first battery of regressions, A refers to the period during treatment, B refers to the period before treatment, and the regressions are run by subgroup (the specified target groups) and by district. In the second battery of regressions, A and B are used to make six pairwise comparisons of each district with each other district during the treatment. In each case, the authors test whether $(\alpha_a + \alpha_b) = \alpha$. The effects of geographic and financial access are captured in the X matrix by distance measures of walking time and income quartiles, respectively. It is unclear from the discussion what the omitted category is in each case. It is also unclear whether the standard errors of the estimates were corrected for the clustered nature of the sample design.

Although the logit techniques are an efficient way of addressing three of the four research questions—utilization patterns, the effect on subgroups, and the effects of geographic and financial access—the fourth question, the effect of changes in cost recovery, is addressed by administrative data and simple comparisons of means. One obvious concern in the latter approach, which was not explicitly addressed, is the possibility of bias in the reporting of the posttreatment results. In particular, there is some moral hazard if administrators are evaluated on the successful response to the treatment.

The effect of the treatments on the use of alternative health systems was addressed through econometric techniques described elsewhere.

V. Who Carried It Out

The Ministry of Public Health carried out the survey with the financial and technical assistance of the U.S. Agency for International Development and the World Bank. The evaluation itself was carried out by Francis Diop, Abode Yazbeck, and Ricardo Bitran of Abt Associates.

VI. Results

The study found that the tax plus fee generated more revenue per capita than the fee-based system, in addition to being much more popular. The tax-based fee system also had better outcomes in terms of providing

access to improved health care for the poor, women, and children. However, because geography is a major barrier to health care access, a tax-based system effectively redistributes the cost of health care from people close to health facilities toward people a long way from such facilities.

The district that implemented fee-for-service saw a slight decline in the number of initial visits but an increase in demand for health care services—compared with a dramatic increase in both in the tax-plus-fee district. Much of this could be attributed to the increase in the quality of the service associated with the quality improvements, which more than offset the increase in cost.

The cost containment—particularly of drug costs—associated with the quality and management reform also proved to be effective and sustainable. Cost recovery in the tax-plus-fee district approached and exceeded 100 percent but was substantially less in the fee-for-service district. In addition, there was much higher willingness to pay in the former than in the latter.

The major result is that the tax-plus-fee approach is both more effective in achieving the stated goals and more popular with the population. The evaluation also demonstrated, however, that lack of geographic access to health care facilities is a major barrier to usage. This suggests that there are some distributional issues associated with going to a tax-plus-fee system: households that are a long way away from health care facilities would implicitly subsidize nearby households.

VII. Lessons Learned

There are a number of useful lessons in this evaluation. One is the multifaceted way in which it assesses the project's impact on multiple dimensions related to sustainability: not only on cost recovery but also on quality and on the reaction of affected target groups. Another is the attention to detail in data collection with both administrative and survey instruments, which then bore fruit through the ability to identify exactly which components of the intervention worked and why. Finally, the analysis of the impact on each target group proved particularly useful for policy recommendations.

VIII. Sources

Diop, F. A Yazbeck, and R. Bitran. 1995. "The Impact of Alternative Cost Recovery Schemes on Access and Equity in Niger." *Health Policy and Planning* 10 (3): 223–40.

Wouters, A. 1995. "Improving Quality through Cost Recovery in Niger." 10 (3): 257–70.

Annex 1.14: Schooling Outcomes in Philippine Elementary Schools: Evaluation of the Impact of Four Experiments

I. Introduction

Project Description. In most developing countries high dropout rates and inadequate student learning in primary education are a matter of concern to policymakers. This is certainly the case in the Philippines: almost one-quarter of Philippine children drop out before completing sixth grade, and those who leave have often mastered less than half of what they have been taught. The government embarked on a Dropout Intervention Program (DIP) in 1990–92 to address these issues. Four experiments were undertaken: provision of multilevel learning materials (MLM), school lunches (SL), and each of these combined with a parent-teacher partnership (PTP). The first approach allows teachers to pace teaching to different student needs and is much less expensive than school feeding. Parent-teacher partnerships cost almost nothing but can help with student learning both at home and at school.

Highlights of Evaluation. The evaluation is noteworthy in that it explicitly aimed to build capacity in the host country so that evaluation would become an integral component of new initiatives, and data requirements would be considered before rather than after future project implementations. However, there are some problems that occur as a consequence, and the evaluation is very clear about what to expect. Another major contribution of the evaluation is the check for robustness of results with different econometric approaches. Finally, the benefit-cost analysis applied at the end is important in that it explicitly recognizes that significant results do not suffice: inexpensive interventions may still be better than expensive ones.

II. Research Questions and Evaluation Design

The key research question is the evaluation of the impact of four different interventions on dropping out and student outcomes. However, the evaluation design is conditioned by pragmatic as well as programmatic needs. The DIP team followed a three-stage school selection process:

- Two districts in each of five regions of the country were identified as a low-income municipality. In one district the treatment choices were

packaged as control, MLM, or MLM-PTP; in the other control, SL, or SL-PTP. The assignment of the two intervention packages was by a coin flip.

- In each district the team selected three schools that (a) had all grades of instruction, with one class per grade; (b) had a high dropout rate; and (c) had no school feeding program in place.
- The three schools in each district were assigned to control or one of the two interventions based on a random drawing.

Each intervention was randomly assigned to all classes in five schools, and both pre- and posttests were administered in both 1991 and 1992 to all classes in all 20 schools as well as in 10 control schools.

III. Data

The data collection procedure is instructive in and of itself. Baseline data collection began in 1990–91, and the interventions were implemented in 1991–92. Detailed information was gathered on 29 schools, on some 180 teachers, and on about 4,000 pupils in each of the two years. Although these questionnaires were very detailed, this turned out to be needless: only a small subset of the information was actually used, which suggests that part of the burden of the evaluation process could usefully be minimized. Pretests and posttests were also administered at the beginning and end of each school year in three subjects: mathematics, Filipino, and English.

The data were structured to be longitudinal on both pupils and schools. Unfortunately the identifiers on the students turned out not to be unique for pupils and schools between the two years. It is worth noting that this was not known a priori and only became obvious after six months of work uncovered internal inconsistencies. The recovery of the original identifiers from the Philippine Department of Education was not possible. Fortunately, the data could be rescued for first graders, which permitted some longitudinal analysis.

IV. Econometric Techniques

The structure of the sampling procedure raised some interesting econometric problems: one set for dropping out and one for test score outcomes. In each case there are two sets of obvious controls: one is the control group of schools, and the other is the baseline survey conducted in the year prior to the intervention. The authors handled these in different ways.

In the analysis of dropping out, it is natural to set up a difference-in-difference approach and compare the change in the mean dropout rate in

each intervention class between the two years with the change in the mean dropout rate for the control classes. However, two issues immediately arose. First, the results, although quite large in size, were only significant for the MLM intervention, possibly owing to small sample size issues. This is not uncommon with this type of procedure and likely to be endemic given the lack of funding for large-scale experiments in a developing-country context. Second, a brief check of whether student characteristics and outcomes were in fact the same across schools in the year prior to the interventions suggested that there were some significant differences in characteristics. These two factors led the authors to check the robustness of the results via logistic regression techniques that controlled for personal characteristics (PC) and family background (FB). The core result was unchanged. However, the regression technique did uncover an important indirect core cause of dropping out, which was poor academic performance. This naturally led to the second set of analysis, which focused on achievement.

A different set of econometric concerns was raised in the evaluation of the impact of the intervention INTER on the academic performance of individual I in school s at time t (AP_{ist}), which the authors model as

$$AP_{ist} = \delta_0 + \delta_1 AP_{ist-1} + \delta_2 PC_i + \delta_3 FB_i + \delta_4 LE_{st} + \delta_5 CC_i + \delta_6 INTER_{jt} + \varepsilon$$

where LE is learning environment and CC is classroom conditions.

First among these issues is accounting for the clustered correlation in errors that is likely to exist for students in the same classes and schools. Second is attempting to capture unobserved heterogeneity. And the third, related, issue is selection bias.

The first issue is dealt with by applying a Huber-White correction to the standard errors. The second could, in principle, be captured at the individual level by using the difference in test scores as an independent variable. However, the authors argue that this is inappropriate because it presupposes that the value of δ_1 is 1, which is not validated by tests. They therefore retain the lagged dependent variable specification, but this raises the next problem—one of endogenous regressor bias. This is handled by instrumenting the pretest score in each subject with the pretest scores in the other subjects. The authors note, however, that the reduction in bias comes at a cost—a reduction in efficiency—and hence report both least squares and instrumental variables results. The authors use both school and teacher fixed effects to control for unobserved heterogeneity in LE and CC.

The third problem is one that is also endemic to the literature and for which there is no fully accepted solution: selection bias. Clearly, because there are differential dropout rates, the individual academic performance

is conditional on the decision not to drop out. Although this problem has often been addressed by the two-stage Heckman procedure, there is a great deal of dissatisfaction with it for three reasons: its sensitivity to the assumption of the normal distribution, the choice and adequacy of the appropriate variables to use in the first stage, and its frequent reliance on identification through the nonlinearity of the first stage. Unfortunately, there is still no consensus about an appropriate alternative. One that has been proposed is by Krueger, who assigns to dropouts their pretest ranking and returns them to the regression. Thus the authors report three sets of results: the simple regression of outcomes against intervention, the Krueger approach, and the Heckman procedure.

V. Who Carried It Out

The data collection was carried out by the Bureau of Elementary Education of the Philippines Department of Education, Culture, and Sports. The analysis was carried out by a World Bank employee and two academic researchers.

VI. Results

The study evaluates the impact of these interventions on dropping out in grades one through six and on test score outcomes in first grade using a difference-in-differences approach, instrumental variable techniques, and the Heckman selection method. The effect of multilevel materials—particularly with a parent-teacher partnership—on dropping out and improving academic performance is robust to different specifications as well as being quite cost-effective. The effect of school lunches was, in general, weak. An interesting component of the study was a cost-benefit analysis—which makes the important point that the story does not end with significant results! In particular, a straightforward calculation of both the direct and indirect (opportunity) costs of the program leads to the conclusion that the MLM approach is both effective and cost-effective.

The lack of effectiveness of school feeding might be overstated, however: it is possible that a more targeted approach for school feeding programs might be appropriate. Furthermore, because there is quite a short period of time between the implementation and the evaluation of the program, the evaluation cannot address the long-term impact of the interventions.

VII. Lessons Learned

Several lessons were learned through this evaluation procedure. One major one was that the devil is in the details—that a lot of vital longitu-

dinal information can be lost if adequate information, such as the uniqueness of identifiers over time, is lost. A second one is that very little of the information that is gathered in detailed surveys was used and that a substantial burden to the respondents could have been reduced. Third, the study highlights the value of different econometric approaches and the advantages of finding consistency across techniques. Fourth, this study is exemplary in its use of cost-benefit analysis—both identifying and valuing the costs of the different interventions. Finally, although errors were clearly made during the study, the authors note that a prime motive for the study was to build evaluation capacity in the Philippines. The fact that the DIP was implemented and evaluated means that such capacity can be nurtured within ministries of education.

VIII. Source

Tan, J. P., J. Lane, and G. Lassibille. 1999. "Schooling Outcomes in Philippine Elementary Schools: Evaluation of the Impact of Four Experiments." *World Bank Economic Review*, September.

Annex 1.15: Assessing the Poverty Impact of Rural Roads Projects in Vietnam

I. Introduction

Project Description. Rural roads are being extensively championed by the World Bank and other donors as instruments for alleviating poverty. The Vietnam Rural Transport Project I was launched in 1997 with funding from the World Bank for implementation over three to five years. The goal of the project is to raise living standards in poor areas by rehabilitating existing roads and bridges and enhancing market access. In each participating province, projects are identified for rehabilitation through least-cost criteria (size of population that will benefit and project cost). However, in an effort to enhance poverty targeting, 20 percent of each province's funds can be set aside for low-density, mountainous areas populated by ethnic minorities where projects would not strictly qualify under least-cost criteria.

Impact Evaluation. Despite a general consensus on the importance of rural roads, there is surprisingly little concrete evidence on the size and nature of the benefits from such infrastructure. The goal of the Vietnam Rural Roads Impact Evaluation is to determine how household welfare is changing in communes that have road project interventions compared with ones that do not. The key issue for the evaluation is to successfully isolate the impact of the road from the myriad of other factors that are changing in present-day rural Vietnam as a result of the ongoing transition to a market economy.

The evaluation began concurrent with project preparation, in early 1997, and is in process. No results are available yet. The evaluation is compelling in that it is one of the first comprehensive attempts to assess the impact of a rural roads project on welfare outcomes—the bottom line in terms of assessing whether projects really do reduce poverty. The design attempts to improve on earlier infrastructure evaluation efforts by combining the following elements: (a) collecting baseline and follow-up survey data, (b) including appropriate controls so that results are robust to unobserved factors that influence both program placement and outcomes, and (c) following the project long enough (through successive data collection rounds) to capture its full welfare impact.

II. Evaluation Design

The design of the Vietnam Rural Roads Impact Evaluation centers on baseline (preintervention) and follow-up (postintervention) survey data

for a sample of project and nonproject communes. Appropriate controls can be identified from among the nonproject communities through matched-comparison techniques. The baseline data allows before-and-after ("reflexive") comparison of welfare indicators in project and control group communities. In theory the control group, selected through matched-comparison techniques, is identical to the project group according to both observed and unobserved characteristics so that resulting outcomes in program communities can be attributed to the project intervention.

III. Data Collection and Analysis Techniques

Data collected for the purposes of the evaluation include commune- and household-level surveys, along with district-, province-, and project-level databases. The baseline and follow-up commune and household surveys were conducted in 1997 and 1999, and third and fourth survey rounds, conducted at two-year intervals, are planned. The survey sample includes 100 project and 100 nonproject communes, located in 6 of the 18 provinces covered by the project. Project communes were selected randomly from lists of all communes with proposed projects in each province. A list was then drawn up of all remaining communes in districts with proposed projects, from which control communes were randomly drawn. (Ideally, controls differ from the project group only insofar as they do not receive an intervention. And for logistical reasons, it was desirable to limit the fieldwork to certain regions. Controls were therefore picked in the vicinity of, and indeed in the same districts as, the treatment communes. Districts are large and contamination from project to nonproject commune is therefore unlikely, but this will need to be carefully checked.) Propensity-score matching techniques based on commune characteristics will be used to test the selection of controls, and any controls with unusual attributes relative to the project communes will be dropped from the sample. A logit model of commune participation in the project will be estimated and used to ensure that the control communes have similar propensity scores (predicted values from the logit model).

The commune database draws on existing administrative data collected annually by the communes covering demographics, land use, and production activities and augmented with a commune-level survey conducted for the purposes of the evaluation. The survey covers general characteristics, infrastructure, employment, sources of livelihood, agriculture, land and other assets, education, health care, development programs, community organizations, commune finance, and prices. These data will be used to construct a number of commune-level indicators of welfare and to test program impacts over time.

The main objective of the household survey is to capture information on household access to various facilities and services and how this changes over time. The household questionnaire was administered to 15 randomly selected households in each commune, covering employment, assets, production and employment activities, education, health, marketing, credit, community activities, access to social security and poverty programs, and transport. Owing to limited surveying capacity in-country, no attempt is made to gather the complex set of data required to generate a household-level indicator of welfare (such as income or consumption). However, a number of questions were included in the survey that replicate questions in the Vietnam Living Standards Survey. Using this and other information on household characteristics common to both surveys, regression techniques will be used to estimate each household's position in the national distribution of welfare. A short district-level database was also prepared to help put the commune-level data in context, including data on population, land use, the economy, and social indicators. Each of these surveys is to be repeated following the commune survey schedule.

Existing information was used to set up two additional databases. An extensive province-level database was established to help understand the selection of the provinces into the project. This database covers all of Vietnam's provinces and has data on a wide number of socioeconomic variables. Finally, a project-level database for each of the project areas surveyed was also constructed in order to control for both the magnitude of the project and its method of implementation in assessing project impact.

The baseline data will be used to model the selection of project sites by focusing on the underlying economic, social, and political economy processes. Later rounds will then be used to understand gains measurable at the commune level, conditional on selection. The analytical approach will be "double differencing" with matching methods. Matching will be used to select ideal controls from among the 100 sampled nonproject communes. Outcomes in the project communes will be compared with those found in the control communes, both before and after the introduction of the road projects. The impact of the program is then identified as the difference between outcomes in the project areas after the program and before it, minus the corresponding outcome difference in the matched control areas. This methodology provides an unbiased estimate of project impacts in the presence of unobserved time-invariant factors that influence both the selection of project areas and outcomes. The results will be enhanced by the fact that the data sets are rich in both outcome indicators and explanatory variables. The outcome indicators to be examined include commune-level agricultural yields, income source diversification, employment opportunities, land use and distribu-

tion, availability of goods, services and facilities, and asset wealth and distribution.

IV. Evaluation Costs and Administration

Costs. The total cost of the evaluation to date is \$222,500, or 3.6 percent of total project costs. This sum includes \$202,500 covering the first two rounds of data collection and a \$20,000 research grant. World Bank staff time and travel expenses are not included in these costs.

Administration. The evaluation was designed by World Bank staff member Dominique van de Walle. An independent consultant with an economics and research background in rural poverty and development was hired to be the in-country supervisor of the study. This consultant has hired and trained the team supervisors, organized all logistics, and supervised all data collection.

V. Source

van de Walle, Dominique. 1999. *Assessing the Poverty Impact of Rural Road Projects*. World Bank, Washington, D.C. Processed.