Annexes

List of Annexes

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 by improving social infrastructure in poor communities. Trabajar 1, a pilot program, was introduced in 1996 in response to a prevailing economic crisis and unemployment rates of over 17%. Trabajar 2 was launched in 1997 as an expanded and reformed version of the pilot program, and Trabajar 3 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 non-government 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 2 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 2 and 3 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 2, and is on-going. The aim of the evaluation is to determine whether or not the program is achieving its policy goals, and to indicate areas where the program requires reform in order to maximize its effectiveness. The evaluation consists of a number of separate studies which 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 foregone 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 spending 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 analyses conducted for a sub-sample 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 2 and Trabajar 3. 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 & 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[1]. 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% 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 2 projects located within

[1] 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.

the EDS sample frame, and generates data for 2,802 current program participants (total Trabajar 2 participants between May '97 and January '98 numbered 65,321). The reliability of the matching technique is enhanced by being able to apply the same questionnaire to both participants and the control group, at the same time, and to ensure that both groups were from the same economic environment.

To generate the matching control group from the EDS survey, the study uses a technique called propensity scoring[2]. 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)[3]. Data on incomes in the matching control group of non-participants allows the income foregone by actual Trabajar 2 participants to be estimated. Net income arising from program participation is then calculated as total program wages minus foregone 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 1 (March 1996-April 1997) and the first six months of Trabajar 2 (May –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' (UBN)[4]. 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 local government.

The cost-benefit analysis was conducted by a civil engineer, who conducted two a two-stage study of Trabajar infrastructure projects. In the first stage she visited a sample of 50 completed Trabajar 1 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

[2] The fact that the EDS questionnaire is very comprehensive, collecting detailed data on household characteristics which help predict program participation, facilitates the use of the propensity scoring technique.

[3] The propensity score is calculated for each observation in the participant and control group sample using standard logit models.

[4] 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 non-working family members).

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 2 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 1 and Trabajar 2. They provide feedback on project implementation issues such as the role of NGOs, 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 foregone income is important to gaining an accurate portrayal of workfare program benefits. Descriptive statistics for Trabajar 2 participants suggest that without access to the program (per capita family income minus program wages) about 85% of program participants would fall in the bottom 20% of the national income distribution – and would therefore be classified as poor in Argentina. However matching-method estimates of foregone income are sizable, so that average net income gained through program participation is about half of the Trabajar wage[5]. However, even allowing for foregone income the distribution of gains is decidedly pro-poor, with 80% of program participants falling in the bottom 20% of the income distribution. Female participation in the program is low (15%), 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 2 reforms. There was a seven-fold increase in the implicit allocation of resources to poor households between Trabajar 1 and Trabajar 2. One-third of this improvement results from better targeting at the central level, while two-thirds results from improved targeting at the provincial level. There are, however, significant differences in targeting outcomes between provinces. A department with 40% of people classified as poor can expect to receive anywhere from zero to five times the mean departmental allocation, depending upon which province it belongs to. Further, these targeting performance tended to be worse in the poorest provinces.

Infrastructure project quality was found to be adequate but Trabajar 2 reforms, disappointingly, did not result in significant improvements. Part of the reason was the sharp expansion of the program, which made if difficult form the program to meet some of the operational standards which 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.

[5] Program participants could not afford to be unemployed in absence of the program, hence some income is foregone through program participation. It is this foregone income which is estimated by observing the incomes of non-participants 'matched' to program participants.

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 results of participation (although these net gains are lower than the gross wage, due to income foregone). The program does not seem to discriminate against female participation. Trabajar 2 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 which merit further policy attention. Finally, disappointing results on infrastructure project quality have generated tremendous efforts by the project team at improving performance in this area by enhancing operating procedures – insisting of more site visits for evaluation and supervision, penalizing agencies with poor performance at project completion, and strengthening the evaluation manual.

VI. Evaluation Costs & Administration

Costs: The costs for carrying out the Trabajar survey (for the study of net income gains) and data processing was approximately $350,000. The two evaluations of sub-project 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 Argentinean project team. Throughout its different stages, the evaluation effort also required coordination with several local government agencies, including the statistical agency, Ministry of Labor (including field offices), and the policy analysis division of the Secretary for Social Development.

VII. Lessons Learned

Importance of accounting for foregone income in assessing the gains to workfare: foregone income represents a sizable proportion (about half) of the gross wage earned by workfare program participants in Argentina. The results 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 using the matched comparison evaluation technique, propensity scores allow reliable matches to be drawn between a participant and non-participant (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 into 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. Source

Jalan, Jyotsna, and Martin Ravallion. 1999. Income Gains from Workfare and their Distribution. The World Bank. Washington, DC. Processed.

Ravallion, Martin. 1999. Monitoring Targeting Performance when Decentralized Allocations to the Poor are Unobserved. The World Bank. Washington, DC. 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 ½ 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% more, have almost twice as high a proportion of sons in school and a substantially increased proportion of daughters compared to the bottom quartile.

Highlights of Evaluation The evaluation investigates the impact of the programs on 1800 households in Bangladesh, and compares them to 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 are 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 program lending has not followed eligibility guidelines – in fact, many of the borrowers had land-holdings in excess of the ½ 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 just 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 micro-finance programs on

i) log consumption per capita
ii) variance of log consumption
iii) log labor per adult in previous month
iv) variance of per adult log labor
v) adult male labor hours in past month
vi) adult female labor hours in past month
vii) % male school enrollment (age 5-17)
viii) % female school enrollment (age 5-17)

The evaluation is survey based, and covers 87 villages surveyed three times during 1991-92. Villages were chosen randomly from a census and administrative lists, from 5 sub-districts that served as controls; and 24 sub-districts 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 ½ acre of land. It is worth reproducing the schematic, because the schematic then illustrates how to create dummy variables that characterize the typology and how to think about selection bias.

<table>
<tr><td colspan="2">Village 1: With program</td><td></td><td>Village 2: Control</td></tr>
<tr><td colspan="2">A
Not eligible [b=1;e=0;c=0]</td><td>Households with more than ½ acre</td><td>B
would not be eligible [b=0;e=0;c=0]</td></tr>
<tr><td>C
eligible but does not participate [b=1;e=1;c=0]</td><td>D
Participants [b=1;e=1;c=1]</td><td>Households with ½ acre and below</td><td>E
Would be eligible [b=0;e=1;c=0]</td></tr>
</table>

Comparing outcomes for group D to 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 to group E, however, because outcome difference depend 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 non-random 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 1798 households; 1538 of these were eligible to participate and 905 actually participate. The surveys were collected in 1991-92 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 non-parametric 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 .5 acre. Finally, the study primarily uses difference and differences 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 Jonathan Morduch of the Economics Department and HID at Harvard University.

VI. Results

The results suggest that almost all the apparent gains from the program are due to selection bias due to loan mistargeting. In particular, the authors find that 20-30% of the borrowers own more land than the ½ acre maximum requirement for the program – suggesting 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 difference 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 perverse after correction for selection bias.

It is also worth noting that while this analysis shows little impact of the treatment relative to the control group, the control group may not, in fact, lacked access to financing since this may be supplied by non-governmental organizations. The expenditure of millions of dollars to subsidize micro finance 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 difference or simple difference techniques. The third is considering the impact of an intervention on the second as well as the first moment of the distribution: since the reduction in risk may, in itself, be a useful outcome. There is a more fundamental lesson, which 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. Source

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

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% in 1993-4 to 26% in 1995-6, and reached 2.2 million children, or 13% 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% 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 – since 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 decision making process. In brief, since 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 to 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 3625 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 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 geographic location helps to model the decision making 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 needs to find 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:

$$W_i = \forall IP_i + \exists' X_i + 0 Z_i + :_i \qquad (1)$$

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, since this is not correlated with the error term, as well as household characteristics. This then characterizes this relationship as:

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

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 is fairly standard, and actually directly affect 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, are worth noting for illustrative purposes. They 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 .55). This suggests that these variables do in fact capture overall placement.

This set of equations can then be modeled using three stage least squares (3SLS), and compared to 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 3SLS approach was 66% higher than the OLS estimates without geographic controls and 49% higher than with the controls. In other words, simple estimates which only control for variation across households (OLS 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% using the method outlined above.

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.

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 which 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 which have since been introduced in Latin American, 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 on-going. The study features baseline (1993) and follow-up (1997) survey data which 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 1992, and is on-going. The design includes separate evaluations of education, health and water projects which 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, since randomization (random selection of program beneficiaries among an eligible group) is widely viewed as the more statistically robust method, and yet matched comparison (using a non-random process to select a control group that most closely 'matches' the characteristics of program beneficiaries) is more widely used in practice.

III. Data Collection & 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 (4155 households, 190 health centers), education projects (1894 households, 156 schools), water projects (1071 households, 18 water projects) and latrine projects (231 households, 15 projects).

The household survey consists of three sub-samples: 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; 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, the study explores whether the characteristics of communities, schools or health centers associated with significantly greater than average impacts can be identified.

In education, where SIF investments were randomly assigned among a larger pool of equally eligible communities, the study applies the 'ideal' randomized experiment design (where the counterfactual can be directly observed). In health and sanitation projects, where projects 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 don't 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 6^{th} 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 vs. 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 using 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, sub-projects 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 sector. 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 to the pre-existing source) have generated much attention, and project design in this sector is being rethought.

VI. Lessons Learned

Effectiveness of 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 where 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 7 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 where there is extensive uncertainty regarding results – where payoffs of the evaluation effort are likely to be greatest.

VII. Evaluation Costs & Administration

Costs: The total estimated cost of the Bolivia SIF evaluation to date is $878,000, which represents 0.5% of total project cost. Data collection represents a relatively high proportion of these costs (69%), 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 (INE), and managed by SIF counterparts for the first round, and by UDAPSO and later the Ministry of Hacienda for the second round.

VIII. Sources

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). Pp. 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 non-participants. 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.

There are several important lessons 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 due to 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 active labor market programs that are the subject of this evaluation. The five ALP's 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 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 re-entering the labor market than are non-participants – 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 non-participant group (with information drawn from administrative records) on people who registered with the state Employment Service, but was not selected for the ALP. The match is more fully described in Box X, but the fundamental notion is that an individual is selected at random from the ALP participant group. This individual's outcomes are then compared with individuals in the non-participant 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 which have been extensively addressed elsewhere (Burtless (1995) and Heckman and Smith (1995). One obvious concern which is endemic to any non-randomized 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 non-participants 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 in 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, not only identified 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 were contacts. Of these, 9,477 participated in ALP in 1994-5. The response rate for non-participants was 14%; for participants it was 24.7%, 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 eight 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 were different, leading to quite different policy implications. Indeed, a major recommendation of the evaluation was the ALP's 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 ALP's 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 which 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 policy makers. 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 9 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 to the increase in earnings and employment.

VII. Lessons Learned

There are several important lessons learned from this study. First among these are the pragmatic components discussed in the introduction - particularly 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 (see Box X). Finally, the evaluation provides a nice illustration of the importance of conducting analysis not just in aggregate but also on subgroups – with the resultant possibility of fruitful targeted interventions.

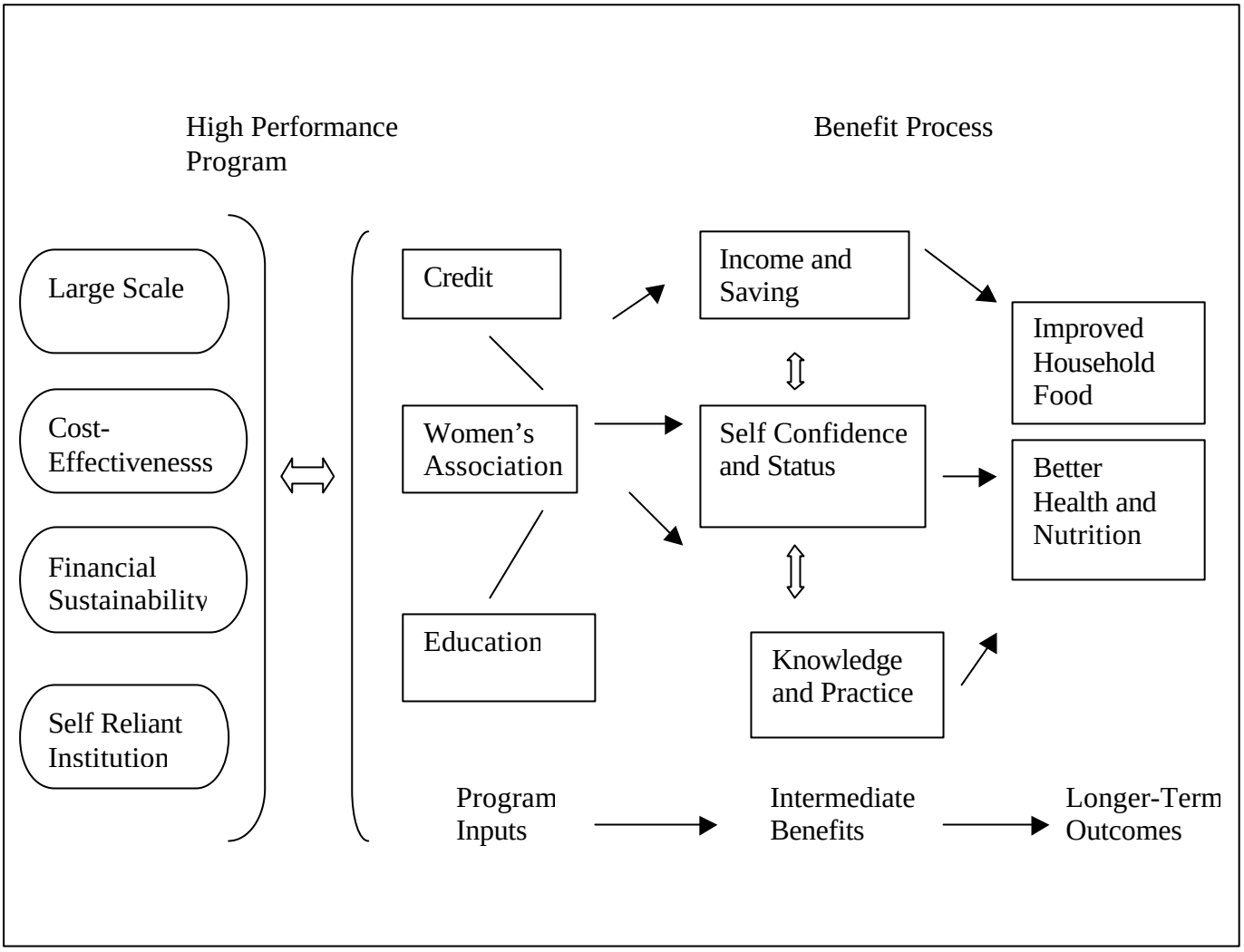
VIII. Source

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

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) Spring 1995 pp. 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 – over 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 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 focussed on the program's effects on:

i) the nutritional status of children
ii) women's economic capacity (income, savings time) to invest in food and health care
iii) women's knowledge and adoption of breast-feeding, weaning, and diarrhea management and prevention practices
iv) 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/nutrition practices. A total of 299 mother/child pairs were surveyed in the first period and 290 different pairs in the second period, gathering both qualitative and quantitative information.

The evaluation design was quite complex. The Lower Pra Rural Bank identified 19 communities which 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 due to 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/child pairs, with children aged 12-23 months, were chosen for the baseline surveys from small communities; thirty 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, since 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 .4 difference in the z-score values of the control and target groups and with a target significance level of .05 and a power of .8.

III. Data

Both quantitative and qualitative data were collected on the household, mother and child, focussing on both intermediate and long-term measures – and particularly the multi-dimensional 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 decision making in the household; 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/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 found no significant differences.

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 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 since, almost by construction, the outcomes of each individual mother/child pair will be correlated with the others in the community, the standard errors will be biased down, and the t-statistics spuriously biased up. . In the extreme case, where 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 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 and correction for this fundamental problem. Finally, although there were significant differences in item non-response rates, suggesting 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/non-participant 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 non-quantitative factors such as women's empowerment. The other key lesson is the value of non-quantitative data to illustrate the validity of quantitative inferences.

VIII. Sources

MkNelly, Barbara and Christopher Dunford (in collaboration with the Program in International Nutrition, University of California, Davis)"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 1998

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 has textbooks in grades three, four and five, rising to one in four in later grades, and 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 3-7 with a ratio of 6 textbooks to every 10 children, math textbooks to grades 3, 5 and 7, with a 50% ratio and science textbooks were provided to grade 8 with a 60% 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 making inference 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. Since this is a complex concept, the outcomes are measured as the difference between textbook and comparison schools in several dimensions

i) post-test scores
ii) test score gains
iii) differences between subject-grade combinations which did and did not receive textbooks

iv) 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 fund raising.

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. Maths, English and Science textbooks were provided to different grades – primarily grades 3-7.

III. Data

Maths, 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 pre-tests that were administered in grades 3-7 in October 1996 and post-tests 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. Since 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 lasted for fifteen 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. 4-5 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 fundraising 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 post test). 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 estimated 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 pre-test scores between control and treatment schools.

The third approach is a little more complicated, because it exploits within school variation, and deserves discussion. The regression applied here involves regressing test scores on dummies that capture whether the students were i) in a textbook school and

ii) in a subject-grade combination that received a textbook (p.10). 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 interacts 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 specifically, 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 arises from differential promotion and repetition rates between textbook and comparison schools. For example, children might be differentially promoted from grade 2 (a non textbook 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 by 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 prevision on parent fundraising. They found that the intervention did crowd out parent contributions, since the amount of non-ICS aid received by comparison schools was $465; 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.

V. Who carried it out

A Dutch non-profit organization, International Christelijk Steunfonds funded the project. The evaluation is performed by an MIT 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

VI. Results

The result of this evaluation was in marked contrast to 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. They also f

VII. 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 sub groups of the population. Finally, it is important to look for and identify secondary effects – in this case, the potential for crowding out.

VIII. Sources

Glewwe, Paul; Michael Kremer; and Sylvie Moulin. 1998. "Textbooks and test scores: evidence from a prospective evaluation in Kenya," processed, Development Research Group (DECRG), World Bank

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 (i.e., farmer awareness and adoption of new techniques) and outputs (e.g., frequency and quality of contact) to assess the performance of the extension system and to confirm the actual, or the potential for, 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.[6] 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

[6] 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.

implementation process. The study also applied the contingent valuation method to elicit farmers' willingness to pay for extension services[7]. 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 pre-project 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 draws on a range of pre-existing data sources in Kenya (household surveys, participatory assessments, etc.), complemented with 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 (OED). This survey generates panel data by revisiting as many households as could be relocated from a 1990 household survey conducted by the Africa Technical Department (ATD), which in turn drew from a sub-sample of the 1982 Rural Household Budget Survey.[8] 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 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

[7] The 'contingent valuation method' elicits individuals' use and non-use 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, i.e., 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.

[8] These three surveys generate a panel data set for approximately 300 households. The surveys cover household demographics, farm characteristics, 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.

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 using the Data Envelopment Analysis (DEA), a non-parametric 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 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 non-existent. The **quality and quantity of service provision** is poor. Beneficiaries and extension service staff alike report that visits are infrequent and ineffective. While 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, indicating 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. Further, no significant impact of the supply of extension services on productivity at the farm level could be established 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 non-recipients, 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 where 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** (M&E) 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 (NGOs) 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, farmer organizations, etc.) 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 & 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 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 (Bank staff).

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 where 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 pre-project 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 with individual methods to 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 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. For More Information

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

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 (Revenga, 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 and the third evaluation 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 received 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 multi-faceted 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. Since 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 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. While 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:

i) the probability of employment after three, six and twelve months
ii) the time to exit unemployment
iii) the effect on monthly earnings, work hours per week and hourly wages

iv) 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 (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. While there is no alternative to this combination of surveys due to data limitations, the construction of the joint sample (control and treatment groups) can be critiqued, as discussed in the 1999 evaluation:

i) in using the unemployed individuals in the ENEU to form the control group, it is assumed that none of the ENEU individuals have benefited form the program. This is not the case since 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 is for 1993-94);

ii) 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 since 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 evaluation are as follows.

i) In the 1994 evaluation, the authors attempt to address the selection bias problems resulting from the PROBECAT's non-random 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.

ii) 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 sixteen main urban areas in Mexico. This survey was part of a regular quarterly labor force survey, ENEU (Encuesta Nacional de Empleo), 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 which became unemployed in the fourth quarter of 1990, and tracked for 9 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:

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

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:

i) 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.

ii) 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 due 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 non zero 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 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 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:

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

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

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 using program availability at the local level as a determinant of individual participation. Then the above equations are estimated 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 independant variable. Again, comparisons of various models show that bias are 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 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, while the last evaluation is not. The disappointing results in the last evaluation are not surprising. Most retraining programs in OECD 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 considered 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. Source

Revenga, Ana, Michelle Riboud and Hong Tan "The Impact of Mexico's Retraining Program on Employment and Wages" World Bank Economic Review, 8(2), 1994 p 247-277.

Wodon Quentin and Minowa, Mari, "Training for the Urban Unemployed: A Reevaluation of Mexico's Probecat", World Bank, Government Programs and Poverty in Mexico, Report No. 19214-ME,Volume 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 Mexio 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 1-1.5 million families at the end of 1998, with an expenditure of 500 million dollars.

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 multi-dimensioned. There are thus many different evaluation components: (1) beneficiary selection; (2) evaluation methods; (3) non-experimental analytical framework; (4) data requirements; (5) impacts on education; (6) impacts on health; (7) impacts on food consumption and nutrition; (8) impacts on consumption expenditures and intra household allocation; (9) potential second-round impacts of the program; (10) simulations of changes in program benefits; and (11) 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 corrective 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, since 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 (1) comparing the distribution of household consumption levels in participant and non-participant households in treatment localities (2) deriving an eligibility cutoff for household consumption which is consistent with the total number of households that PROGRESA can serve (3) conduct sensitivity and specificity analysis of PROGRESA and non-PROGRESA households versus the households selected and not selected under this cutoff (4) explore the ability of current criteria to predict consumption (5) identify alternative criteria from other data sources and (6) simulate models which could improve targeting with alternative criteria (IFPRI Proposal for Evaluation, p.6)

For the impact evaluation, the same system was followed, with the result that localities were randomly allocated to 296 treatment and 173 non-treatment 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 non-randomization by comparing the characteristics of treatment and control groups. If they are systematically different, then three non-experimental 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 stakeholder. The main focus is on identifying what and why things are happening, the level of satisfaction 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 non eligible 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 sub sample 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 consists of screening, targeted delivery mechanisms and monitoring costs; the latter includes foregone income generation.

IV. Econometric Techniques

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

$$W_j = \alpha_I + \beta_{1j} \text{ lpcexp} + \beta_{2j} \text{ lsiz} + \Sigma_k \delta_{kj} \text{ dem}_k + \Sigma_s \Theta_{sj} z_s + \beta_{3j} P + e_j$$

Where: w_j is the budget share of the jth good; lpcexp 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 survey allows difference in difference methodologies to be used.

They also identify key econometric issues which 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:

i) policy changes may be hard to predict because of cross-substitution and behavior adjustment
ii) marginal benefits and marginal costs depend on a number of things: externalities (putting a wedge between social and private valuation); the actors (parents vs. children)
iii) the importance of unobserved characteristics
iv) the importance of controlling for individual, family and community characteristics
v) the empirical estimates depend on a given macro economic, market, policy and regulatory environment.

VII. Source

International Food Policy Research Institute, Programa Nacional de Educacion, Salud, Y Alimentacion (Progresa): A Proposal for Evaluation (plus technical appendix), May 1998

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 which 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 is expected to be almost universal by end-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 Educational Reforms represents one of the first systematic efforts to evaluate the impact school decentralization on student outcomes. The evaluation, carried out jointly by the World Bank and the Ministry of Education, began in 1995 and will be complete by end-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 decision-making processes. The evaluation also illustrates 'best practice' techniques when there is no baseline data, and when selective (non-random) 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), while the quantitative results suggest that increased decision-making by schools is in fact significantly associated with improved student performance.

II. Evaluation Design

The design of the Nicaraguan Education Reform evaluation is based on a technique called 'matched comparison', where data for a representative sample of schools participating in the reform process is compared with data from a sample of non-participating schools. The sample of non-participating schools is chosen to most closely as possible 'match' 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 (where the sample of schools studied in the evaluation would be random, and therefore nationally representative).

III. Data Collection & Analysis Techniques

The qualitative study draws on data for a sample of 12 schools, 9 reformers and 3 non-reformers which represent the control group[9]. The sample of 12 schools was picked to represent both primary and secondary schools, rural and urban schools and, using data from the 1995 quantitative survey, with differing degrees of actual autonomy in decision-making. 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 2-4 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 which was conducted in two rounds (Nov.-Dec. 1995, and Apr.-Aug. 1997), and student achievement tests for students in these schools which were conducted in Nov. 1996. The school survey collected data on school enrollment, repetition and dropout rates, physical and human resources, school decision-making, and characteristics of school director, teachers, students and their families. The school decision-making module is unique, and presents a series of 25 questions designed to gauge whether and how the reform has actually increased decision-making by schools. The survey covered 116 secondary schools (73 reformers and 43 non-reformers representing the control group), and 126 primary schools (80 reformers and 46 non-reformers). 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, with 10-15 students chosen at random from each school. Those students that

[9]]Data was actually gathered for 18 schools, but only 12 of these schools were included in the qualitative study given 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.

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, while 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 parent participation, as well as management and leadership improvements. But the degree of success with which reforms are implemented varies with school context. Of particular 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 (where key actors such as teachers do not feel integrated into the reform process, success at decentralization has been limited). Policy makers 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 decision-making, nor does it 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[10]. 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 increasing the number of textbooks, teacher training, class size, etc.

[10] This result is preliminary pending further exploration using the panel data, which has recently come available.

V. Policy Application

The evaluation results provide concrete evidence that Nicaragua's School 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 decision-making does result in improved student achievement.

The results also point out areas where 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 in 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 & Administration

Costs: The total cost of the evaluation was approximately $495,000, representing less than 1.5% of the World Bank credit[11]. Of this total evaluation cost, 39% was spent on technical support provided by outside consultants, 35% on data collection, 18% on World Bank staff time, and 8% 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 *(title*), Nora Gordon *(title*) and Nora Mayorga de Caldera (*title*). 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 while others had failed, and hence help guide policy adjustments. Furthermore, because it is more intuitive, the qualitative work was more accessible and therefore interesting to Ministry staff, which in

[11]This total does not include the cost of local counterpart teams in the Nicaraguan Ministry of Education.

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 into the reform, as it was evolving. The policy impact of the evaluation was also enhanced by a cohesive local team in which evaluators and policy-makers worked collaboratively, and because the Minister of Education was brought on board as an integral supporter of the evaluation process.

VIII. Source

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. The World Bank. Washington, DC.

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. The World Bank. Washington, DC.

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

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. The World Bank. Washington, DC.

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. The World Bank. Washington, DC.

Rawlings, Laura. Forthcoming. *Assessing Educational Management and Quality in Nicaragua*, in [*book title*]. The World Bank. Washington, DC.

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 – making it important to effectively allocate those resources. 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 didn't. All classes received both a pre test at the beginning of the year and a post test at the end. The study then used simple regression techniques to compare the mean classroom post-test scores as a function of pre-test 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 pre-existing 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 focussing 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 to radio instruction as well as to 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" (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 piggy-backed on a pre-existing project evaluation. In the existing Radio Nicaragua Project, an entire mechanism has 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% of students are in rural schools, but equal numbers of classes were chosen for in each stratum).

The textbook evaluation exploited this pre-existing 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/urban). They then randomly assigned schools to treatment or control from these master lists for each category, and then schools 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. 20 of the control schools received textbooks for each child, and teachers received both written and oral instruction and the teacher's editions of the tests. The radio component consisted of 150 daily mathematics lessons, combined with student worksheet and written and oral teacher instructions.

An interesting decision that was made is the deliberate lack of supervision of treatment groups. This was clearly difficult, because the absence of supervision made it difficult 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, were intended to be given in the last three weeks of school, but were administered two weeks early because of political problems. The students had, as much as possible, identical conditions for test taken: both because they had the same length of time for the test 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 pre-test/post-test comparisons. Although there were a total of 20 control classes, 20 textbook classes, and 47 radio classes, the numbers of pre-test and post-test scores were different in each group because of late registration, dropping out, absence and failure to be tested due to overcrowding. Individual information on the students does not appear to have been collected.

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

i) there were no standardized test in use in Nicaragua;
ii) The test had to assess the achievement of the curriculum objectives;
iii) 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; ¼ 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 post-test scores for treatment and control groups are derived by adding average scores for each test, and the standard errors are calculated 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 twenty 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 post-test score was regressed against the mean pre-test 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 post-test scores by 3.5 items correct; radio lessons by 14.9 items – quite substantial given that the classroom standard deviation is 8.3 and 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 pre-test scores.

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

VII. Lessons Learned

There are three main lessons 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, which 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 "Improving Elementary Mathematics Education in Nicaragua: An Experimental Study of the Impact of Textbooks and Radio on Achievement" Journal of Educational Psychology, 73(4), August 1981556-567

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 (non-hospital) sector in Niger. Niger is a poor, rural economy; public health costs are 5-6% of the government budget; and much of this financing is mistargeted towards 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 to examine 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 was based on 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 for different target groups (children, women, village 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:

i) the demand for and utilization of public health care facilities

ii) specific target groups (poor, women, and children)

iii) financial and geographic access

iv) the use of alternative services

v) 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 each were similar in terms of economic, demographic and social characteristics, they are 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 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 $ .66) for a user over 5, a fee of 100 FCFA for a user under 5. The second combined an annual tax of 200 FCFA paid by district taxpayers and a fee of 50 FCFA per user over 5 and 25 FCFA for children under 5. Annual income was under $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 1800 households. The baseline survey had information on 2833 individuals who had been sick the two weeks before the survey and 1770 childbearing women; the final survey had data on 2710 sick individuals and 1615 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 (P_I) to specify the following model:

$$\text{Logit}(P_I) = X\exists + * (A+B)$$

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

$$\text{Logit}(P_I) = X\exists + *_a A + *_b B$$

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 6 pairwise comparisons of each district to each other district during the treatment. In each case, the authors test whether $(*_a + *_b) = *$. The effect 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; the effect 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 post-treatment 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 USAID and the World Bank. The evaluation itself was carried out by Francis Dip, 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, since 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 towards people a long way from such facilities.

The district which 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 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%, but was substantially less in the fee for service district. In addition, there was much higher willingness to pay in the former than 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. It 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 project's impact on multiple dimensions related to sustainability: not only 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 "The impact of alternative cost recovery schemes on access and equity in Niger" Health Policy and Planning, 10(3) 223-240

Wouters, A. "Improving quality through cost recovery in Niger" Health Policy and Planning 10(3) 257-270

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 policy makers. This is certainly the case in the Philippines: almost ¼ 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 multi-level 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 will 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:

i) 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, SF of SF-PTP. The assignment of the two intervention packages was by a coin flip

ii) In each district the team selected three schools which: a) had all grades of instruction, with one class per grade b) had a high drop out rate c) no school feeding program was in place

iii) 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 post tests administered to in both 1991 and 1992 to all classes in all 20 schools, as well as in 10 control school

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 1991-2. 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 – suggesting that part of the burden of the evaluation process could usefully be minimized. Pre-tests and post-tests 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, permitting 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, 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, which was possibly due 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 focussed 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 (APist), which the authors model as:

$$AP_{ist} = \delta_0 + \delta_1 AP_{ist-1}, + \delta_2 PC_i + \delta 3 FB_i + \delta_4 LE_{st} + \delta_5 CC_i + \delta 6 INTER_{jt} + \varepsilon$$

First among these is accounting for the clustered correlation in errors that is likely to exist for students in the same classes and schools. The 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 this 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 pre-test score in each subject with the pre-test 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 learning environment (LE) and classroom conditions (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, since 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 this 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 1-6 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 multi-level 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 – making 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 lead 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, since there is quite a short period of time between the implementation and evaluation of the program, the evaluation cannot address the long term impact of the interventions.

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 longitudinal 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 DIP was implemented and evaluated means that such capacity can be nurtured within ministries of education.

VII. Source

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

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

I. Introduction

Project Description: Rural roads are being extensively championed by the World Bank and other donors as poverty alleviation instruments. The Viet Nam Rural Transport Project I (RTPI) was launched in 1997 with funding from the World Bank for implementation over 3 to 5 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% 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 Viet Nam Rural Road Impact Evaluation is to determine how household welfare is changing in communes that have road project interventions compared to 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 Viet Nam 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 upon earlier infrastructure evaluation efforts by combining the following elements: 1) collection of baseline and follow-up survey data; 2) including appropriate controls, so that results are robust to unobserved factors influencing both program placement and outcomes; and 3) following the project long enough (through successive data collection rounds) to capture its full welfare impact.

II. Evaluation Design

The design of the Viet Nam Rural Road impact evaluation centers on baseline (pre-intervention) and follow-up (post-intervention) survey data for a sample of project and non-project communes. Appropriate controls can be identified from among the non-project 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 & 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 non-project 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[12]. 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[13].

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. Due 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 Viet Nam Living Standards Survey (VNLSS). 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,

[12] Ideally, controls differ from the project group only in so far 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 non-project commune therefore unlikely, but this will need to be carefully checked.

[13] A logit model of commune participation in the project will be estimated, and used to assure that the control communes have similar propensity scores (predicted values from the logit model).

land use, the economy, social indicators, etc. Each of these surveys is to be repeated following the commune survey schedule.

Two additional databases were set up using existing information. An extensive **province-level database** was established to help understand the selection of the provinces into the project. This database covers all of Viet Nam's provinces and has data on a wide number of socio-economic variables. Finally, a **project-level database** for each of the project areas surveyed was also constructed, in order to control both for 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 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 of 'double differencing' with matching methods. Matching will be used to select ideal controls from among the one hundred sampled non-project communes. Outcomes in the project communes will be compared to 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 influencing 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 distribution, availability of goods, services and facilities, and asset wealth and distribution.

IV. Evaluation Costs & Administration

Costs: The total cost of the evaluation to date is $222,500, or 3.6% 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. The World Bank. Processed.

Annex 2: Sample Terms of Reference

Example I. The Uganda Nutrition and Early Childhood Development Project[14]

Terms of Reference for Consulting Firm to assist in the Project Evaluation

I. Background

The Government of Uganda (GOU) has applied for a credit from the International Development Association (IDA) towards the cost of a Nutrition and Early Childhood Project. The project focuses on improving the quality of life of children under six (6) years of age and building the capacity of families and communities to care for Children. Specifically, the project will aim at achieving early child development through improving Nutrition, Health, Psycho-social and Cognitive status of children under six years of age in Uganda.

II. Rationale for investing in Early Childhood Development

Investing in Early Childhood Development (ECD) has tangible benefits for not only the children and parents but also for entire communities and the country. Rapid physical growth and mental development occur during infancy and early childhood such that at two years of age, a child's brain is nearly fully grown. Cognitive abilities are also developed to a large extent by four years of age. Adequate physical and mental growth and development during early childhood enhances school readiness, improves school retention, and contributes to human capital dependency. Children from disadvantaged backgrounds can particularly benefit from early child-care thus bridging the gaps and inequalities associated with poverty.

Good health and nutrition are crucial as is mental stimulation, if the child is to develop secure conceptual structures in later life. The synergy between nutrition, health and mental stimulation is so crucial that tangible positive effects on child growth and development can only be achieved through an integrated approach.

[14] These terms of reference were prepared by Harold Alderman

III. Project Objectives and Strategies

The development objective of the project is to improve growth and development of children under six years of age, in terms of nutrition, health, psycho-social and cognitive aspects. The achievement of these objectives at the end of the five year implementation period will be measured by the following markers: (a) reduced prevalence of underweight pre-school children by one-third of the 1995 levels in the project districts; (b) reduced prevalence of stunting on entry into primary schools by one-fourth of the 1995 level in the project districts; (c) improved children's psycho-social and cognitive development; (d) reduced repetition and drop-outs at lower primary school level (d) development of entrepreneurship skills and economic empowerment of mothers and caregivers.

The project supports the Ugandan National Program of Action for Children, and the Poverty Eradication Action Plan. The project particularly enhances school readiness of young children and thus contributes towards reaching the goal of Universal Primary Education. The main project strategy is to enhance the capacity of families and communities to take better care of pre-school age children (0 to 6 years) through enhancing knowledge on child growth and development, parenting, nutrition and health care, and income generating activities for women.

IV. Project Approach

The project is a process-driven locally prioritised program rather than a blueprint package. Inputs are to be phased into communities as a result of a participatory planning process to ensure ownership and sustainability. The program will involve collaboration between government and non-government entities(including local and international NGOs) and communities. As a multi-sectoral program involving health, nutrition, early childhood education, child care, savings and income generation, the approach will involve linking various government departments and the non-government entities to provide a comprehensive service towards development of children. The project will support a range of options, a program menu, relating to the needs of pre-school children and their families.

V. Project Components

Project Component 1 - Integrated Community Child Care Interventions

This component supports the Government's goals (a) to improve parental awareness on major aspects of child care, growth and development through parental education, child growth monitoring and promotion, training and sensitization; and (b) to empower communities to support child development programs through capacity building, through skills for income generation and through support grants. The objective is to reduce malnutrition (low weight for age) of children by a third at the end of the five-year period in the project districts, and increase readiness of children for primary schooling

and thereby contribute to the drive for Universal Primary Education. The government plan is to eventually cover all districts, however, interventions in this phase will be implemented in 25 districts chosen by the government based on the level of malnutrition, infant mortality, and rate of primary school enrolment. The project includes the following interrelated interventions:

(a) Parental Education. This sub-component will increase parents' and caregivers' understanding of major aspects of child care, growth and development including child nutrition, health, cognitive and psycho-social development. A range of related competencies will be strengthened in parents. Building parental skills and knowledge will in turn improve health, psycho-social development and well-being of children and, ultimately, their receptiveness to education at the primary level. The program will mobilize groups of mothers (and parents) at the community level, supported by project materials in local languages, technical supervision and communications. Simplified learning materials for adults with low literacy have been tested successfully in Uganda. Emphasis will be on the enhancement of **child care practices** that promote proper growth and development of children, including: childhood nutrition and health (**exclusive breastfeeding and appropriate weaning practices**--particularly the period of introduction of weaning foods, as well as the type of foods given, and food preparation, **child growth promotion, and deworming), psycho-social development, cognitive stimulation and social support, hygiene and improved home health practices.**

The above interventions shall be strengthened/ supported by an outreach activity (children's day) organized at Parish level to enable communities access a number of child related services at a one stop shopping. A study of the impact of providing the anathelminth albendazole to young children in selected parishes will also be conducted in the course of parish-based child days and will measure the effect of six-monthly treatments on weight gain.

(b) Community Capacity Building and Empowerment for Child Care. This sub component comprises two interrelated activities: (i) community capacity building conducted through community planning and sensitization workshops (ii) training in entrepreneurship to increase incomes of mothers and caregivers.

Project Component 2 - Community Support Grants for Child Development.

Two types of grants would be available to communities:

(a) Community Support Grants to communities are offered on the basis of matching contributions from communities. These grants and contributions from communities will cover activities designed to support interventions for child development and which fall within the guidelines and menu contained in the Project Implementation Manual. To qualify for this grant, communities will provide counterpart contributions which maybe in the form of goods, works or services. Examples of the uses of such grants are: construction and operation of community child care centers, home based child

care centers, or the production and marketing of weaning foods. The support grants component will be implemented in the same 25 districts in Component 1.

(b) Innovation Grants are grants made available to communities to address child related problems. The Innovation Grant will provide grants to implement interventions outside the menu of interventions described by the community support grants (a) above. As the term implies, the "innovation" fund will be used to support communities at different levels in implementing "innovative ideas" on improving the lives of children within their communities. The Innovation Grant will be accessed by communities in the same manner as the community support grants: that is, proposals prepared by communities following a participatory planning exercise, followed by screening by a sub-county committee, and forwarded for funding by the project.

Project Component 3 - National Support Program for Child Development

This component consists of central program activities and policy initiatives designed to support the district level programs in components 1 and 2, and provide quality assurance for the frontline project activities at the community level. This component includes:

(a) program monitoring and evaluation,
(b) support for prevention of micro nutrient deficiencies,
(c) early childhood development (ECD) curriculum development,
(d) training of trainers for ECD,
(e) information, education and communications (IEC).

VI. Implementation arrangements

The implementation of the project is the responsibility of Government of Uganda (GOU) assisted by Non Government Organizations within the decentralization framework and devolution of powers to lower levels as stipulated in national policies. The community (LC-1) is the unit of operation for service delivery although the co-ordination structure will also involve the Parish (LC-2), the Sub county (LC-3) and the District (LC-5) levels.

In addition, the project hopes to use stakeholder sensitization and consultations, community mobilization, participatory community planning, capacity building for sustainability at all levels, together with strengthening of complementarity with existing national programs and structures. Use of existing political and institutional structures should be made use of in a multi- sectoral manner. Transparency and accountability should also be ensured at all levels.

VII. Project Co-ordination

National

A Project steering Committee (PSC) composed of line ministries, donors and NGO representatives will be responsible for overall guidance of project implementation.

The committee, to be headed by the Permanent Secretary in the Ministry of Finance, Planning and Economic Development (MFPED) will provide guidance to the project on policy issues, review and approve quality and efficiency of implementation. The PSC will also make suggestions to improve the district annual budget and work plans for the project.

A small project coordinating office (PCO) composed of a coordinator, a deputy coordinator, a qualified accountant and a small support staff, will be based in the social services sector of the MFPED and will take responsibility for the day to day co-ordination of project activities at the national level.

District.

An existing multi-sectoral committee will be identified by the CAO (District Coordinating Committee (DCC)) to take on the responsibility of coordinating the project at the district level. The CAO will identify a focal person from among the government officers who will co-ordinate the NGO related and other activities in the project. The Lead NGO will be included as a member of the DCC. In districts where no NGO with adequate/ appropriate capacity and skill base can be identified or strengthened to take over as the lead NGO, the implementation will be through the district administration.

Sub-county.

An existing sectoral committee similar to the one at the district level will be responsible for coordinating the project activities at the sub-county. This Sub County Coordination Committee (SCC) will also facilitate linkages between existing structures and those of the project and along with the lead NGO for the district approve the sub-county NGO/CBOs annual work plans and funding requirements for the project in the sub-county.

VIII. Project Impact Evaluation

The GOU is interested in assessing the impact of various aspect of the project in order to ascertain its effectiveness and to guide the design of further NECD project. Moreover, as the World Bank considers this project to potentially inform other countries regarding NECD services, it has included the project in a three country evaluation of ECD programs and will provide technical assistance on a grant basis to the PCO to assist

specific research activities. In particular, two studies to evaluate the impact of specific project interventions will be undertaken as part of the overall project:

1) ***Parish Child Health Day Study*** for assessing the coverage of anthelmintic treatments given at parish level child health days and their impact on the weight gain of children under 6 using a randomized experimental design.

2) ***Survey Research using baseline and resurvey methodology*** for assessing:

2.1 the impact of anthelmintic treatments and of overall project activities on the cognitive development, health and nutrition of children under 6 years of age

2.2 the impact of the caregiver education component and mass media communication campaign in the knowledge, attitude and child rearing practices of the principal caregivers.

2.3 the impact of grassroots management training, income generating activities and credit savings group formation, and provision of community grants in household and community welfare.

The Firm will provide technical and logistical support for the above studies and will be invited to participate as local research implementers in the design, data collection, and analysis necessary to complete the two studies of impact assessment. This Firm will be the primary counterpart of the PCO, local researchers, and the researchers from the World Bank and the University of Oxford who will be undertaking the impact assessment.

IX. Overview of Studies

Study One: Impact of Deworming at Parish child days

There have been a number of studies indicating the impact of treating school aged children with anthelmintic medicine. However, there is only one large-scale, randomized trial that shows large effect on weight gain for pre-school aged children. This has raised the question of whether such an effect could be achieved in African children. Thus, the NECD project will include a randomized study of the impact of providing the anthelminth albendazole to young children in 25 selected parishes in the course of parish-based child days and to measure the effect of six-monthly treatments on weight gain. Data will be collected from these parishes as well as 25 control groups which will also organize child health day but will not administer albendazole on a routine basis. If the anthelmintic treatments are delivered successfully and are shown to have beneficial effects on Ugandan children, then the program of anthelmintic treatment may be recommended for all Districts.

Because this is a scientific controlled trial, the selection of parishes which will be asked to administer albendazole will be undertaken by the PCO from a list of parishes where child days will be organized (this list will be provided by the NGOs working in the districts). The PCO will also select parishes which will serve as the control group. This experimental design is key to a successful evaluation.

The Firm will ensure that the local NGOs responsible for the organization of the child health days in the parishes are aware of the rationale for the experimental design and that they comply with the strategy. Each child aged 12 months or older and less than 6 years who attends the fair in the 25 designated parishes will be given a single 400 mg tablet of chewable, proprietary albendazole. The albendazole will be administered every 6 months; in the event that the NGOs chose to organize child days on a more frequent basis the anthelmintic will *still* be administered on a six month schedule and not more often.

Children in parishes where albendazole is administered *as well as* children in the 25 designated control parishes will be weighed at each child day and their weights recorded both on their own health card and on the community register. Children who are too small to stand on the scale unaided will be weighed in their mother's arms after the scale has been set to zero with the mother standing alone on the scale. These weights will be recorded to the nearest tenth (0.1) of a kilogram. The data on the community registers is the responsibility of the local NGOs, although the Firm will work with the NGOs to ensure that the data collection system is compatible with the full range of objectives of the study.

The Firm consultant will be transcribe these weights on a proforma to be designed with technical advisors from the World Bank and the University of Oxford. This data transcription will be undertaken every six months after the child day in the project area. In addition to the child's ID (a unique combination of the parish ID, the village ID and the individual ID recorded on both the child's own card and the community register) the data on the proforma will include the child's gender, the date of birth of the child taken from the child's health card or if not available the age of the child taken from the parish register, the date of the child fair at which the weights were recorded, and whether or not the child took a dose of albendazole. This data will be entered in a computerized record in Kampala. The individual ID will provide the basis to merge the data from different periods and, thus, the ID must be recorded *each* time the data is transcribed and must remain the *constant* for a child over the entire project.

The local circumstances and conditions at each child day which may deter mothers from attending will also be recorded. This includes data on the state of the harvest and the weather conditions, both of which may deter mothers from attending. Any special methods and opportunities used to advertise each child day will be recorded because different forms of advertising may affect attendance. The record should also include an estimate of the number of children who visited the child day from other parishes who did not have ID numbers obtained from the organizers of the child day.

The experiment will last two years. Thus, the Firm consultant will record this data five times for each parish. That is, the Firm consultant will collect the data at the beginning of the project and at 6, 12, 18, and 24 months after project initiation.

A complete copy of the data will be sent to the PCO every 6 months. These copies of the data will be considered the deliverable services of the first study of the project. Preliminary analysis will be undertaken at the University of Oxford on a semi-annual basis. However, the Firm is requested to nominate a representative who will participate in the main analysis to be performed at the end of two years. This representative will be provided travel and living expenses to work on the analysis at Oxford. The funds for this travel are budgeted in a separate line item and therefore need *not* be included in the contract covered by these RFP.

Study Two: Overall Impact of NECD Interventions

Household surveys and community surveys will collect baseline and follow-up information needed to evaluate the impact of the various project activities. The surveys will have several modules which will measure:

(i) cognitive development and growth of children under 6 years of age resulting from anthelmintic treatments and of overall project activities;

> Study two will assess longitudinal growth and psycho-social and cognitive development outcomes in a cohort of children in communities participating in the project (with and without anthelmintic treatment) compared with a cohort of children in non-participating communities. Both cohorts will be followed for two or more years. The study will, therefore, complement the study of deworming at the parish levels by allowing a greater understanding of the decision to take children to child days and to measure whether, over time, participation leads to an increase in measures of cognitive development. Moreover, by including communities which do not receive any ECD services, the study will assess whether the *package* of services leads to improvements in nutritional status and cognitive development.

(ii) changes of knowledge, attitude, and child rearing practices of the caregivers resulting from project parental education and mass media campaign;

(iii) improvement of the health and nutrition of children under 6 years of age resulting from growth monitoring activities, preventive health and nutrition education, anthelmintic treatments, and overall project activities;

(iv) household welfare resulting from community grants, grassroots management training, income generating activities and credit savings group formation.

(v) community characteristics and changes resulting from the project interventions (or otherwise) which could have an impact on child well-being during the duration of the project.

Sample selection

The basis for this study will be a baseline survey collected at the time services are first delivered to the communities and a follow-up survey collected from the same households two years after the initial survey. One third of the sample will be drawn from the same 25 parishes in the treatment (anthelmintic) group and another third from the control groups studies in study one. In addition, one third of the sample will come from villages in 25 parishes in the same districts as the treatment groups but which are not expected to receive services from the NECD project. Thirty (30) households will be selected from each parish. This implies 750 households per strata (2250 total) in the initial survey. Given expected sample attrition, 5-10% fewer households are expected in the resurvey.

To collect the sample in the treatment and control parishes, all households in each parish (there are approximately 700 households in a parish on average) will be listed, possibly by a resident of the community. This list will contain the name of the household head, an indication of the location of the household, and the number of children under 6 in the household. This list will serve two purposes. First, a sample of 30 households containing at least one child under the age of 6 *per parish* will be selected by a random draw. Second, the total number of children under six will serve as an estimate of the potential coverage of children in child days and thus, assist in determining the rate of attendance.

Since the NECD project will have less contact with the communities which have no current NECD activity, the selection of households which receive no ECD service should use cluster sampling to reduce the costs of sample listing. In particular, one subcounty which is *not* in the project should be selected for every subcounty that is in the treatment group, preferable one that is adjacent to it. All parishes in these subcounties should be listed and a random draw of 25 parishes from the total list will be selected. Two villages from each parish selected will then be chosen, again using a list of all the villages in the parish. This step reduces the number of villages where a census will be need to be conducted. The census -similar to the one used in the treatment and control parishes - will form the list of households used to draw the sample of 30 households per parish. This will be the third strata of the survey.

The initial baseline survey should be undertaken in mid 1999. This timing is based on the need to know the sub-counties and parishes in which the NGOs will be operating in order to employ the suggested sample design. This timing is also based on the assumption that the selection and training of lead NGOs will not be completed until late 1998.

The development and pre-testing of the questionnaire, however, should be undertaken much earlier than this (early 1999) in order to be ready to implement the survey as soon as the NGOs have identified the parishes in which they will be working. As the baseline needs to be fielded before the first deworming, the ideal time for the baseline survey is concurrent with the initial community organization that will lead to a child day. Since the sample of 30 families in each parish is small relative to the total population, it is unlikely that the survey data collection will disrupt other activities or overburden the communities. The data collection in the control groups (those with NGO activity but no deworming and those with neither) should be simultaneous with the data collection in the treatment group.

Survey Instruments

The basic questionnaires to be used for the survey project are: household questionnaires (which gather data at the level of the household and individuals) and community questionnaires.

X. Household Survey:

Household data collected using a precoded schedule. This will be drafted on the model of the Living Standards Surveys used in over 30 countries. A first draft will be provided by researchers from the World Bank. However, the instrument will both be abridged to accommodate the particular needs of the project and adapted to local conditions using focus groups and a pretest procedure undertaken by the Firm. The household questionnaire will contain modules to collect data on:

1) Socio-demographic characteristics: a roster of individuals residing in the household in the last 12 months, their age and gender, as well as their schooling and type of employment (if any). The coding format will indicate the parents of all children, if present, and if not present, whether the parents are still living. A detailed list of assets will be collected to serve as an indicator of socio-economic status.

2) *Knowledge, attitude and practices:* The questionnaire will also collect information on the knowledge, attitudes, and child rearing practices of the principal care givers.

3) *Anthropometric data:* Weights will be recorded to the nearest tenth (0.1) of a kilogram for all children under the age of 6 using digital scales to be provided. In addition, heights will be collected for all children between the age of 2 and 6. The pretest will be used to determine if it is feasible to collect the weights of the mothers of these children (if living in the households) as well.

4) *Cognitive assessment: The firm will work with other local and international research consultants to the PCO to integrate tests of child cognitive development into the overall field data collection.* In the baseline survey an internationally recognized test of cognitive development will be administered to children aged 4.0-5.99 years. This test

will also be administered to the same age group in the second round of the survey, allowing a comparison of cohorts. In addition, a subset of children aged 6-7.99 at the time of the second round will be administered this test. [Annex table summarizes this strategy].

In addition, knowledge assessments based on specific content from the program, and a dynamic assessment may be developed for the second round of the survey. The inclusion of these measures will be evaluated during the course of the project. Finally, a school performance measure will be developed for assessing knowledge acquired in the first year of school and administered to a subset of older children in the resurvey. Existing tests might be adapted.

5) Child health: morbidity data (including number and kind of symptoms, levels of severity, length in time), patterns of access to and utilization of health services, sanitation, etc.

6) Household economy. *The best approach to collecting this information will be extensively explored in the pretest phase and assessed jointly with advisors from the PCO prior to finalizing the questionnaire. The variables may include:* food expenditures, agro-pastoral activities, consumption of home production, non-food expenditures, housing characteristics, inventory of durable goods, employment, economic activities, income, land, crops and animals, income from project activities, household enterprises, and asset ownership. Credit and saving information on amount of money and goods lent and borrowed, if money and goods have been borrowed in the last 12 months, savings and net debt the day of the interview, information on loans, including the schedule, reason for borrowing, and number of loans from the same source, location of savings, if any, including bank, housing savings bank, rural savings bank, etc. This information will be part of the baseline and final surveys only.

XI. Community Survey:

Community questionnaires will be used to gather information on local conditions that are common to all households in the area. *The best approach to collecting this information will be extensively explored in the pretest phase and assessed jointly with advisors from the PCO prior to finalizing the questionnaire. The variables may include:*

1) Demographic information: number of households, total population, population under 6, ethnic groups and religions;

2) Economic information including principal economic activities and patterns of migration for jobs.

3) Infrastructure: access to roads, electricity, pipe water, market, bank, and public transport. Condition of local infrastructure such as roads, sources or fuel and water, availability of electricity, and means of communications.

4) *Local agricultural conditions and practices:* type of crops grown in the community, how often and when they are planted and harvested, and how the harvest is generally sold and qualitative data on rainfall, climate conditions and seasonality.

5) *Education:* Number and types of pre-schools, formal and informal ECD arrangements, distance to schools, number of classes, enrollment rates (gross and by gender), attendance, grade progression, health and nutrition services provided at school (e.g school health programs, school lunch).

6) *Health:* type of health facility and distance and travel time to the nearest of each of several types of health facilities (hospital, pharmacy, health post, etc). Types distance and travel time to the nearest of each of several types of health workers (doctor, nurse, pharmacist, midwife, community health worker, etc).

7) *Other:* number and type of active local NGOs/CBOs, other child related projects or interventions (e.g. government vaccination campaigns), and other community development projects.

Suggested Survey Staff:

Core Survey Staff: composed of the survey manager, the field manager, data manager, and data entry staff will be responsible for overall field supervision, coordination, and monitoring of data collection and data entry and data management activities.

Field Survey Staff: The field operations will be conducted by teams composed of a supervisor, two (or three) interviewers responsible for the main questionnaire and the anthropometric measurements, and a driver. A similar number of specialists in administering tests of cognitive development to the children will be selected and trained in collaboration with local and international experts.

Coordinator for the randomized trial: The coordinator will assist in the development of the data collection instruments, training of local NGOs responsible for the organization of the child days in the parishes on rationale for the experimental design, data collection and data transcription. He/she will be oversee data entry and management of the study data set and will participate in the main analysis to be performed at the end of study.

Organization of field work

The Firm will participate in the drafting of the field instruments prior to the pre-testing of the survey and will have primary responsibility for the pretest. After the pretest the questionnaire will be redesigned (in partnership with researchers from the World Bank) and then translated into local languages.

The Firm will work with other local and international research consultants selected by the PCO to integrate tests of child cognitive development into the overall field data collection. The local ECD researcher, assisted by international consultants, will select and adapt the principal cognitive test to be used and will train the testers.

The following organization of field work is suggested. This is based on international experience and designed to ensure quality control. Some variation of this approach might be agreed upon in consultation with researchers from the World Bank based on the experience of the Firm and other advisors to the PCO and information gained during the pretest.

The field work will be organized into small teams consisting of a supervisor, two (or three) interviewers responsible for the main questionnaire and the anthropometric measurements, and a similar number of specialists in administering tests of cognitive development to the children. These staff will be trained in Kampala by the local ECD researcher in coordination with international advisors on psychological testing. The training will include a discussion of the research objectives, a review of each step of the interview, practice training in the office, a dry run in the field, and a recap of experience after this dry run.

Once teams are trained they should be retained for the entire round of the survey, if possible. However, since a few staff may prove to be unsuitable during the field work, it is advisable to train a few extra staff . It is not advisable to hired hire staff to work for a few days only in one parish and then new staff in the next parish as this results in inexperience staff. All staff should receive new training at the beginning of the resurvey.

During the administration of the cognitive test, children should, to the degree possible, be alone with the interviewer when they are tested. In no case should another person (adult or child) respond to the questions asked of the child. However, during the resurvey the test for the subset of 8 year old test may be administered in a group format if it proves convenient.

The supervisor will be responsible to ensure that the interviewers undertake the survey in the households chosen for the sample without substitution and that all children in the appropriate age groups are administered the tests of cognitive development. In addition, the supervisor will review each questionnaire after completion (prior to the team moving to a new parish) and to ensure that there are no gaps in the questionnaire or to see that seemingly inconsistent information be verified.

The Firm will enter all survey data as soon after it is collected as possible. Copies of the household and child specific data and rating scales along with the documentation necessary to access the data will be provided to the PCO in a computerized format at the end of the baseline survey. The original questionnaires should be retained by the Firm since it is generally the case that the original data will need to be accessed during the course of the analysis.

The child level data must contain accurate identification codes which can be *matched* with the household survey codes. While the unique individual and household codes provided to the PCO need not contain the *names* of the households or their exact location, this information must be stored by the Firm in a manner that makes it possible to revisit the household at a later date. As one step in the analysis will link individuals in the resurvey with their test results from the baseline all individual and household codes must be *held constant* over the three surveys.

XII. Specific Tasks For Survey Specialists

The Firm will participate in the following activities in collaboration with the PCO, local researchers, and the researchers from the World Bank and the University of Oxford, and implementing NGOs:

- Revision of work programs

•Development/adaptation of the data collection instruments and support documentation, including: listing materials, questionnaires, coding guides, interviewer and supervisor manuals, manual of operations, data entry manual, and field procedures.

•Revision of various drafts of documents, layout, translation, back-translation and field testing. Provide hard copies and electronic versions of all documentation to PCO. Forward questionnaires to the World Bank researchers for their review and revision prior to the pilot test.

•Dwelling Listing and Cartographic Updating. The responsibilities for listing of households/dwellings in each selected parish include: obtaining base maps, preparing listing materials, contacting local officials to inform about listing operation, identifying boundaries, drawing maps, listing households in a systematic manner, obtaining household preliminary information on household, including name of the household head, an indication of the location of the household, and the number of children under 6 in the household. Documenting procedures at the time of the sample design, at the end of the fieldwork and at the completion of the data file.

•Preparation of sampling framework (with sampling specialist), training of staff to implement the designed sample, supervision of the implementation stage to ensure the quality of the sample selected, and provision of detailed report outlining all the steps involved in the design and implementation of the sample.

•In consultancy with the World Bank, participate in determination of an appropriate strategy for identifying comparison groups (ie non-project parishes).

•Selection and training of field workers. This activity consists of all the work necessary to develop training materials and manuals for all persons involved in field

work. Training will be required for: (i) interviewers; (ii) supervisors of interviewers; (iii) supervisors of teams; (iv) data entry personnel; and (v) anthropometric personnel.

•Field operation including logistical arrangements for data collection, and obtaining household and individual consent. Keeping a study household register.

•Production of Progress Reports: The Firm will prepare field work progress reports (at six month intervals) copied to the PCO and the World Bank. The Firm should also prepare a basic description of the survey. This should include the survey content, the sample plan and its implementation and the field work techniques used. A full questionnaire and basic documentation should be included as appendices.

•Development of a data entry program using software able to check for ranges and consistency of the data and generate reports indicating missing data, data outside of the accepted ranges and inconsistent answers.

•Data cleaning, data entry, database management and tabulation plans: including development of data entry program, data entry manual, data entry operator training, data quality checks, and guidelines for using the data. Also, coding open-ended questions, verification of the data, checking anthropometric data against standard reference tables.

•Enforcing Data Use Policy Agreement*: The Firm and researchers involved in the process of data collection and analysis will sign a memorandum of understanding with the PCO which will explicitly state the policy regarding issues such as access to data, intended users, procedures for obtaining copies of the data sets and documentation, and publication and authorship rules.*

• Conducting Data analysis: *The Firm will conduct exploratory data analyses (e.g. frequencies, percentages tabulation and cross-tabulations) of key survey variables and its correlates. The Firm will conduct modern statistical modeling of impacts after rounds 2 and 3 to determine overall progress in social indicators (e.g. nutrition, health, incomes, community development) and the factors which account for the changes or lack of changes.*

• Producing Analyses Reports: The firm will report on the findings after rounds 2 and 3 of the surveys based on the analyses of the social indicators and the co-variantes. The Firm will coordinate with the PCO and the World Bank on the Parish Child Health Day Study and on the collection of impact on cognitive development, but will NOT be responsible for the final reports on the result of these studies.

Specific Tasks for Community Survey:

•*Work with advisors from the PCO in the development of the Community Questionnaire and extensively explore in the pretest phase the best approach to collecting this information.*

•*Work closely with the implementing agencies (Lead and Local NGOs) in the collection of the community data.*

•Contact local officials and community leaders to explain the project impact evaluation approach and obtain communal consent for survey research and child health day study.

•Interview key informers, obtain maps, lists and other community records.

•Obtain list of health and education facilities (pre- and primary schools), including geographical location, catchment area, type of establishment (e.g. private, public),

•Obtain community demographic information, including number of households, population by gender and age

•*Obtain other data required in the community questionnaires.*

Specific Tasks for Child Day Study:

•Participate in the development of study protocol

•Development of data collection instruments

•Training of local NGOs responsible for the organization of the child days in the parishes on rationale for the experimental design.

•Supervision of data collected during child day

•Data transcription

•Data entry and management

•Participate in the main analysis to be performed at the end of study

Proposed sample sizes for impact evaluation of Nutrition and Early Child Development Project, Uganda

CATEGORY	Deworming and Parent Education		No deworming and Parent Education		No deworming and no parent education		
No. Parishes	25		25		25		TOTAL
Time	Baseline	Second Round[15]	Baseline	Second Round	Baseline	Second Round	
No. households	750	700	750	700	750	700	2250
No. children weighted at Child Days[16]	5000	5000	5000	5000			20000
No. children with anthropometry in the home ages 0-5.99 (mean 2 per family)[17]	1500	1395	1500	1395	1500	1395	11580
No. Children given cog. Tests: test all children ages 4.0-5.99 in households	500[18]	465[19]	500	465	500	465	2895
No. Children given cognitive test and anthropometry aged 6.0-7.99		subset[20]	-	subset	-	subset	subset
School enrollment rates	25 comm	25 comm	25 comm	25 comm	25 comm	25 comm	

Validity study. In addition to the above, one small longitudinal study will be added to examine the predictive validity of the preschool measure for school performance at the end of the first year of school. In the baseline survey, 2 children per community aged 6.0 t0 6.9 (not yet in school) will be tested, for an N=150. These children will be located at the posttest, and given a school performance test two years later, aged 8.0 to 8.99.

Task schedule.

[15] Assuming a small loss to attrition of 8% in two years

[16] Assuming that about 200 children will attend each Child Day

[17] Two children per family are assumed, but families will be recruited if they have ANY children under 6. Family refers here to a mother (or substitute) child pair.

[18] This is a maximum; the actual number can vary according to the number of 4-5 year old children encountered.

[19] Assuming the same loss of 8% over two years; only children whose parents were interviewed will be tested.

[20] Number will be a subset of the children in this age range whose parents were interviewed. They will be linked with the earlier score. Even though the number of children tested increases in the second round, the time for the interviews may decrease, as much information will not need to be assessed again. It is also possible that the size of this group will be reduced.

Tentative timetable:

Month 1 - Begin process of constructing indicators of cognitive development in conjunction with international consultant and in accord with TOR. This process may take up to six months.

Month 2 - Initial pretest and revision of questionnaire.

Month 5 - Begin listing of households for sample selection. This step is dependent on the selection of the lead and local NGOs. It cannot be done until the PCO and NGOs chose the parishes where child days will be organize and then select the sites for the initial deworming program. At the same time the questionnaire should be translated and field tested again.

Month 7. Begin collection of data at child fairs for the deworming study. Data will be collected at these fairs at 6 month intervals. As above, the timing of this step is dependent on the selection of the lead and local NGOs.

Month 8. Training of field staff for household survey and initiation of survey. The survey should take approximately 3-4 months depending on the number of teams employed. Data entry should be concurrent with data collection.

Month 14. Initial analysis of baseline data. This will be an ongoing process.

Month 20. Staff from Firm visit to University of Oxford to participate in analysis of initial data.

Month 20 - 36. Collection of data for round 2 for deworming study.

Midterm and Final Household Surveys will be conducted 2 and 4 years after baseline.

Support to Firm

No specific support will be given to the firm to carry out assignments. Firms are advised to include all requirement for effective carrying out of the assignment in their proposals.

Example II: Rural Roads Impact Evaluation: Viet Nam 1997 baseline[21]

Terms of Reference: Baseline Survey For Rural Roads Impact Study

I. Background

The study aims to assess the impact on living standards of the World Bank financed Viet Nam rural transport project which will be implemented in 15 poor provinces over 3 to 5 years starting in 1997. The study's overall focus will be on how the determinants of living standards are changing over time in communes which have road project interventions as compared to ones that don't. This requires the collection of pre-project baseline data for both project ("treatment") areas and non-treatment control areas and a number of further data collection rounds of post-intervention data at two yearly intervals. A detailed commune level data base will be created in part by drawing on annually collected records at the commune level. The latter will be augmented by the collection of retrospective commune level data and the collection of various other key supplementary data. A short district level survey will help put the commune level data in context. Finally 10 to 15 households will be randomly sampled from commune level household lists and a short household questionnaire administered. The study will be conducted in 6 provinces out of the 15 that will benefit from the project. The 6 provinces will be representative of 6 geographical regions of Viet Nam. A random sample of around 200 or so project and non-project communes will be drawn. Six teams will be set up to simultaneously survey each province. The survey should begin in April and finish around August. Data should be available around October/November.

II. Survey Design

Sampling: 1. Provinces: The 15 project provinces are located in Viet Nam's 6 geographical regions. Criteria for selection of survey provinces will be the following: a) one province will be selected in each geographical region; b) when there are more than one possible project province in each region, a random selection will be made.

2. Communes: The aim is to survey 200 or more communes, which are randomly selected. About half or less (not more) should be communes with road link projects, the rest controls. A list will be drawn of non-project communes in the 6 provinces (or alternatively one list for each province) and a random sample drawn. Similarly, a list will be drawn of all communes benefiting from road projects in the 6 provinces (or by province). This may be more than one commune per road link; all will be included in the sampling frame. Of these a random sample will also be drawn. The sample will not necessarily include both communes linked by a road project. If access to certain sampled communes is impossible, it will be replaced with another commune in the district which is similar.

[21] These terms of reference were prepared by Dominique van de Walle.

3. Households: In each sampled commune, a household survey will be administered to 15 households. These (plus perhaps a few replacement households) will be randomly selected using the commune household lists. After selection, the commune authorities will be asked about where the households fall in the very poor, poor, average, not poor, rich classifications.

III. Survey Process

Six survey experts will be hired to conduct the surveys in the 6 provinces. After their training and the field testing of the questionnaire, they will begin surveying simultaneously in each province. In the districts, surveyors will need at least one local staff from the District Project Management Unit to help with contacting local authorities and in some cases to help find suitable guides and interpreters in minority areas. Survey assistants or assistance from the Provincial Project Management Units will be hired as required.

Each surveyor will collect data from 35 communes on average, the districts they belong to and 15 or so households per commune. Three to four days will be needed for each commune. The time spent in the field will be around 100 to 140 days (4 to 5 months). The total time will be 6 months.

During the survey period, the supervisor will conduct field visits to all 6 provinces to supervise data collection and ensure high quality.

The collected data will be cleaned and entered using a data entry program.

The following table gives an estimated time table for the study:

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
Design questionnaires	****	****							
Field test survey		****							
Revise questionnaires			****						
Adapt data entry program, translate & print questionnaires			****						
Hiring & training surveyors			****						
Survey in field				****	****	****	****	****	
Data checking								****	****
Data entry									*****

IV. Other

Equipment: The equipment purchased under the project will belong to the project as long as the study continues (through future rounds) but when not in use by the team will be housed in PMU18 for their use.

Budget disbursements: The budget for the study (excluding payments to the main investigator who will receive monthly installments) will be disbursed in three installments. The first, upon signing of the contract will consist of 20% of total funds. The second installment consisting of 50 % of the total budget will be disbursed once the commune, household and district questionnaires have been finalized and approved by the World Bank task manager. This is expected to be sometime in late March. The third and final installment will be disbursed in late July or half way through data collection.

Estimated Study Budget

	No.	Time	Unit amount	Total
1. Main investigator	1	9 mths	$1000	$9000
2. Survey experts	6	6 mths	$400	$14,400
3. Travel allowance for 6 surveyors, 6 local guides/interpreters	12	125 days	$8	$12,000
4. Car & other transport for 6 survey teams	6	125 days	$40	$30,000
Car rental for main investigator	1	30 days	$50	$1500
5. Airtickets Hanoi-Hochiminh-Hanoi - For surveyors(south provinces) - For main investigator	6 3 persons 3 trips		$200	$1200
6. Training of surveyors - payment - travel to field - allowance	12	 1 week 3 days/ 3 cars 3 days	 $50 $50 $8	$1338
7. Field test of questionnaire (South and North communes)	1	2 weeks		$2000
8. Data cleaning and entry	2	2 mth	$200	$800
9. Survey materials				$2000
10. Communications(fax, phone, email, zerox)				$2000
11. Equipment - Computer (PMU18) - printer (PMU18) - fax machine (study team)	 1 1 1 1		 $1700 $1000 $500 $1800	$5000

- laptop computer (study team)				
12. Translation (questionnaire, manuals, documentation)	200 pages		$8/page	$1,600
13. Printing, zeroxing				$800
14. Contingencies				$1362
Total				$85,000

Terms of Reference: Survey Supervisor/Main Investigator

I. Study Background

The study aims to assess the impact on living standards of the World Bank financed Viet Nam rural transport project which will be implemented in 15 poor provinces over 3 to 5 years starting in 1997. The study's overall focus will be on how the determinants of living standards are changing over time in communes which have road project interventions as compared to ones that don't. This requires the collection of pre-project baseline data for both project ("treatment") areas and non-treatment control areas and a number of further data collection rounds of post-intervention data at two yearly intervals. A detailed commune level data base will be created in part by drawing on annually collected records at the commune level. The latter will be augmented by the collection of retrospective commune level data and the collection of various other key supplementary data. A short district level survey will help put the commune level data in context. Finally 10 to 15 households will be randomly sampled from commune level household lists and a short household questionnaire administered. The study will be conducted in 6 provinces out of the 15 that will benefit from the project. The 6 provinces will be representative of 6 geographical regions of Viet Nam. A random sample of around 200 or so project and non-project communes will be drawn. Six teams will be set up to simultaneously survey each province. The survey should begin in April and finish around August. Data should be available around October/November.

II. Job Description

The in-country survey supervisor/main investigator will be responsible for the study's baseline survey work within Viet Nam. Responsibilities include: determining availability of information at the commune level; helping to revise and finalize the district, commune and household level questionnaires; field testing the questionnaire; incorporating revisions to the questionnaire; arranging for the questionnaire to be translated; hiring and training the assistants; planning the field work logistics; preparing survey implementation and questionnaire documentation; supervising survey implementation and ensuring quality control; supervising the project data base and

arranging for data cleaning and entry. The person will also act as the liaison with the Ministry of Transport's PMU18, the World Bank resident mission, the CIDA representative in Hanoi and the Project Task Manager at the World Bank in Washington. The person will report directly to the task manager. The person will start as soon in January 1997 as the contract can be processed for a period of 9 months at a rate of $1000 per month.

III. Specific tasks include:

1. Responsibility for hiring, drafting detailed terms of reference, training and supervising 6 main assistants who will work with local assistants (possibly from the local transport office) in the field and will be responsible for the collection of the district, commune, and household level data.

2. Exploration of data availability at the commune level and working closely with the World Bank task manager to design the final versions of the questionnaires.

3. Carrying out a field test of the questionnaires in both South and North rural communes; reporting back on potential problems and necessary revisions; revising the questionnaires where needed.

4. Arranging for the questionnaires to be translated, printed and xeroxed. The final versions of the questionnaires will be available in both English and Vietnamese.

5. Choosing the 6 provinces to be included in the survey such that there is one province to represent each geographical region. When there are more than one such province, the sampled province is chosen randomly. Drawing up a random sample of around 200 rural communes in the 6 provinces including about half with projects and the rest without projects.

6. Planning all field work logistics including arranging for transport, drivers, travel allowances, the schedule of commune surveys, alerting commune administrations of team arrivals and purpose.

7. Participating in survey implementation, alternating between the teams in a supervisory role. Ensuring quality control. Identifying problems affecting survey implementation, checking quality and completeness of data collected, suggesting ways to solve problems and implementing them following consultation with the study leader.

8. Ensuring that future survey rounds can replicate the baseline survey. This requires the preparation of i) detailed documentation of all survey implementation design and logistics: how the sampling of provinces, communes, and household was done; how the survey teams were trained and organized; how field work was organized; what was the procedure followed when a sampled site was not accessible or a sampled household not found; problems, issues raised, solutions found. Preparing ii) detailed manual on definitions of terms (eg unemployment, income, primary occupation, child/adult, distance

etc), units, currency amounts, codes, used in the questionnaires; how the questionnaires are to be administered--to whom; how prices were collected etc. The former should ensure that future rounds of the survey can reproduce the baseline's organization and logistical details. The latter should be used in training of surveyors and for their work, as well as to

aid future users of the data. There will be both English and Vietnamese versions.

9. Procuring the necessary equipment as itemized in the study budget.

10. Establishing good relations and ensuring close cooperation with PMU18. Keeping them abreast of the study and monitoring project developments. Supervise the setting up of a data base of project specific data. The World Bank task manager will identify the data to be included.

11. Arrange and supervise data cleaning and entry using the provided data entry program.

12. Liaison and communicate often with the task manager.

Annex 3: A sample budget from an Impact Evaluation of a School Feeding Program:

Phase I: July 1999-December 2000*
School Feeding Research Proposal – Baseline and Cross-Sectional Evaluation
(July 1999 - December 2000)
Draft Budget - 7/14/99 –US$

	Staff Weeks/Activity		Source of Funds/Costs			
	FY00	FY01	BB	RPO	Other	Total
World Bank Staff						
Economist	4	2	17,640			
Evaluation Specialist	5	3	23,520			
Nutrition Specialist	5	3	23,520			
Peer Reviewer	0.2	0.2	1,948			
Peer Reviewer	0.2	0.2	1,948			
						68,577
FES Staff						
Study Coordinator	4	4			12,000	
						12,000
International Consultants						
Situational Assessment (incl. Travel)					7,000	
Cognitive Test Development (incl. Travel)				6,000		
Sampling Specialist				2,000		
Cost-Effectiveness Study				25,000		
						40,000
Regional Consulting Firm**						
Design, Sampling, Administration				42,000		
Fieldwork				25,000		
Data Processing				3,500		
Analysis				30,000		
						100,500
Travel to Country						
Trips	4	2		12,000		
						12,000
Contingencies						
Communication				1,000		
Software				2,000		
Translation				2,000		
						5,000
TOTALS			68,577	150,500	19,000	238,077

Total Requested from RAD: $150,500
Total Requested from Bank Budget: $68,577
Total Provided by Outside Sources: $19,000

* budget estimates for phase II of the evaluation are not included in this proposal
** a breakdown of these costs is provided on the next page

Estimated Budget -- Local Data Collection and Analysis for Phase I
School Feeding Impact Evaluation -
Costs in US$

	# People	# Staff Weeks	Weekly Rate	Total,$
Professionals				
Director	*1*	*12*	*2000*	*24,000*
Education Specialist	*1*	*8*	*1500*	*12,000*
Nutrition Specialist	*1*	*8*	*1500*	*12,000*
Statistician/Sampling	*1*	*12*	*750*	*9,000*
Fieldwork Manager	*1*	*8*	*750*	*6,000*
Programmer	*1*	*10*	*300*	*3,000*
Data Processing Supervisor	*1*	*8*	*300*	*2,400*
Assistant – Surveys	*1*	*10*	*100*	*1,000*
Assistant – Anthropometrics	*1*	*10*	*100*	*1,000*
Assistant - Cognitive Tests	*1*	*10*	*100*	*1,000*
Data Quality Control	*1*	*8*	*100*	*800*
Sub Total - Professional staff				72,200
Field Work-Staff				
Supervisor	*4*	*6*	*200*	*4800*
Cognitive Tester	*4*	*6*	*120*	*2880*
Anthropometrist	*4*	*6*	*120*	*2880*
Interviewer	*4*	*6*	*120*	*2880*
Driver	*4*	*5*	*100*	*2000*
FIELD WORK-EQUIPMENT	**People/Units**		**Cost per Week or Unit**	
Vehicles (4 vehicles for 5 weeks)	4	5	350	*7000*
Gasoline (4 vehicles for 5 weeks)	4	5	80	*1600*
Scales; rulers (5 sets)	5		20	100
Cognitive Test Equipment (for 4 testers)	4		20	80
Survey Equipment (for 4 interviewers)	4		20	80
Subtotal - Fieldwork				24300
Data Processing	**People**			
Data Coding	*3*	*7*	*75*	*1575*
Data Entry	*4*	*7*	*75*	*2100*
Sub total - data processing				3675
Total				**100,175**

Annex 4: Impact Indicators, Evaluation of Bolivia Social Investment Fund

Developed November 1997

I. Formal Education--Schools Type "A" And "B" (multigrade and regular)

1. Final Impact Indicators

Achievement in Mathematics and Language tests*
Repetition rate
Dropout rate
Enrollment
Instruction level
Demand for education (% of students rejected from school)*

2. Intermediate Impact Indicators

Regularity in student attendance
Regularity in teacher attendance
Students' time allocation/hours spent studying
Classroom teaching method*
Turnover in teaching staff*

3. Intervention Indicators

Infrastructure
- Ratio m2/student
- Ratio students/classroom
- Number of classrooms in "good shape"
- Number of missing classrooms
- Availability of multifunctional area
- Availability of basic services
  - Electricity
  - Source of main water supply
  - Type of sanitation service; condition of the sanitation service

Furniture
- Ratio students/desk
- Ratio teacher's tables/classroom
- Ratio teacher's chairs/classroom
- Ratio "adequate blackboards"/classroom
- Ratio shelves/classrooms

Texts and didactic material
- Ratio texts/student
- Quality of mathematics, language, social studies and natural sciences texts

* not considered in baseline

Availability of teachers' texts
Availability and condition of maps and chart
Didactic games by school cycle (prebasic, basic and intermediate)
Availability of an abacus

Education Reform Indicators[22]

3. Factors Affecting Outcomes Not Linked To The Sif Project (Exogenous)

Nutrition
Availability of school breakfast program
Cost of the school
Teachers' characteristics
Educational background
Years of service
Training received
Methods applied in teaching (in a period of classes)
Training received, by topic and course
Student evaluation practices (frequency of homework and its correction)
Evaluation of the teachers by the students
Rationale for dropout
Students rejected by the school
Distance between the house and the school
Ratio students/teacher

4. Identification Indicators

Whether school was prioritized by the Education Reform
Programmed cost by project component
Actual expenditures by project component

[22] to be developed in coordination with Education Reform staff ; will be considered as exogenous to the intervention unless the SIF-Education Reform interventions are considered jointly

II. Health

1. Final Impact Indicators[23]

Infant mortality rate
Childhood mortality rate
Rates of incidence and prevalence of main diseases (EDA and IRA)
Prevalence of malnutrition (general, slight, moderate and severe)

2. Intermediate Impact Indicators

Use of government (MSSP) health centers
Prevalence of tetanus vaccination
    Place where vaccine was received
Prevalence of prenatal control
    Number of prenatal controls
    Quality of control
Prevalence of births attended in health centers
    Quality of attention
Prevalence of home births attended by medical personnel
Height at birth
Weight at birth
Anthropometric assessments
    Place where assessment is held
    Age when first assessment is made
Incidence of disease and prevalence of immunization by number of doses received
    Polio
    DPT
    Measles
    BCG
Knowledge of places where to go for immunization
Incidence and treatment for coughing
Incidence and treatment for diarrhea
Prevalence the knowledge and use of oral rehydration packets
Clinics' knowledge of prevalence of pregnancy
Attendance of high risk pregnancies
Prevalence of good hygiene habits and use of water
Duration of lactation

3. Intervention Indicators[24]

Quality of infrastructure by type of health center
Availability of basic services in the health center (drinking water, sewage system and electricity).

---

[23] general mortality rate, birthrate, global fertility rate, adult mortality and life expectancy at birth deleted

[24] training in health topics deleted

Adequacy of infrastructure based on established norms by type of health center
Adequacy of equipment based on established norms by type of health center
Number of beds in the health center
Availability of essential medicines by type of health center
Availability of essential medical instruments by type of health center
Availability of essential furniture by type of health center

4. Factors Affecting Outcomes Not Linked To The Sif Project (Exogenous)

Characteristics of the household
Quality of household
Type of household
Basic Services in the household
Electricity
Source of water
Type of sanitary service
Accessibility to basic services
Distance between the household and the closest health center
Distance between the sanitary service and the source of water
Distance between the household and the main source of water
Hours of availability of water per day
Sufficiency of amount of water per day
Availability of water throughout the year
Cost of consultation in the health center
Household head's perception of the quality of:
- the "service" in the health center attended by the household
- the "infrastructure" of the health center attended by the household
- the "availability of medicines" in the health center attended by the household

Household expenses
Personal characteristics of the members of the household
Age
Language
Education level
Occupation
Geographic characteristics
Health district
Health area
Health sector
Province
Locality
Human resources in the health center (doctors, odontologist, nutritionists, nurses, nurses' [i]assistants, technicians, administrative staff)
Population under the influence area of the health center by age groups
Cost of consultation in the health center

Health interventions not financed by the SIF

5. Identification Indicators

Programmed cost by project component
Actual expenditures by project component

III. Water

1. Final Impact Indicators[25]

Infant mortality rate
Childhood mortality rate
Rates of incidence and prevalence of diarrhea in households
Prevalence of malnutrition (general, slight, moderate and severe)

2. Intermediate Impact Indicators

Incidence and treatment for diarrhea in health centers
Prevalence of use and knowledge of use of oral rehydration packets
Prevalence of good hygiene habits and use of water

3. Intervention Indicators (Of Input)

Prevalence of training in health topics
Accessibility to basic services
- Main supplying source of water
- Existence of sanitary service. Type of sanitary service
- Distance between the sanitary service and the source of water
- Distance between the household and the main source of water
- Hours of availability of water per day
- Sufficiency of amount of water per day
- Availability of water throughout the year

Quantity of water consumed by the household*
Quality of water*

4. Factors Affecting Outcomes Not Linked To The Sif Project (Exogenous)

Use of government (MSSP) health centers
Size at birth
Weight at of birth
Duration of lactation
Characteristics of the household
- Quality of household
- Type of household

Accessibility to basic services
- Distance between the household and the closest health center

[25] general mortality rate, birthrate, global fertility rate, adult mortality (male and female), life expectancy at birth, prevalence of acute respiratory infections and treatment of coughing deleted

* not considered in the baseline

Cost of consultation in the health center
Household's expenses

Personal characteristics of the household's members

Age
Language
Education level
Occupation

5. Identification Indicators

Programmed cost by project component
Actual expenditures by project component

Annex 5: Template of Log Frame for Project Design Summary

Hierarchy of Objectives	**Key Performance Indicators**	**Monitoring and Evaluation**	**Critical Assumptions**
Sector-related CAS Goal: Provide a one-sentence statement of the long-term strategic goal (as reflected in the CAS) to which the project is designed to contribute. The statement should describe substantive development change in the sector(s) of interest.	**Sector Indicators:** 1. Indicators accompanying the sector-related CAS goal involve measurements that are not generally funded by the project, but that may be funded by the Bank as part of other work. 2. Normally the borrower would monitor these indicators as part of good practice sectoral management.	**Sector / Country Reports:** 1. This column identifies where the information for verifying each indicator will be found, and the process involved. 2. Indicators accompanying the sector-related CAS goal are generally monitored and/or evaluated via various sector or country reports generated outside the project.	**(from Goal to Bank Mission)** • Assuming that the sector-related CAS goal (stated in the far left box) is achieved in the long term, list any additional assumptions needed to link this goal to the Bank's mission (i.e., poverty alleviation). • These assumptions often involve conditions, actions, or responses outside of the project and outside of the sector.
Project Development Objective: 1. Provide a one-sentence statement of the behavioral change expected from the target beneficiary-group or institution(s) by the end of project implementation. Achievement of the objective serves as a simple test of demand for project outputs. The objective should express a single development purpose that is realistic, specific, measurable, and demand-driven. For a guide to setting the project objective, see "Do's and Don'ts for Setting a Project Development Objective" (call x37065 or e-mail M&EHelp@worldbank.	**Outcome / Impact Indicators:** 1. Outcome indicators relate to the results to be achieved by the end of project implementation, while impact may not be fully achieved until five or more years after project implementation has been completed. 2. Indicators at the outcome (PDO-level) are not a restatement of those at the output level. 3. Collection of data for measurement of these indicators is generally funded by the project.	**Project Reports:** 1. This column identifies where the information for verifying each indicator will be found, and the process involved. 2. Indicators accompanying the project development objective are generally monitored and / or evaluated via various project reports, supervision mission reports, and evaluation (mid-term and final) reports. 3. Where data collection is required, specific mention should be made of methods and responsibilities,	**(from Project Development Objective to Sector-related CAS Goal)** • Assuming that the project development objective is achieved, list any additional assumptions needed to justify the project's contribution to the stated goal. • These assumptions refer to the contribution(s) of additional projects, additional inputs, or additional responses from beneficiary groups and institutions that are critical to the achievement of the stated goal.

org for a copy)		which may include inquiries from beneficiaries.	

Output from each Component: 1. State here (in the past tense) the value added by the completion of each component. 2. A correct statement of output value added will be easy to measure (as reflected in the indicators to the right). 3. For simplicity and clarity of the logic, there should be one output statement for each corresponding project component. 4. Each output should correspond in number to its respective component. 5. The project team is generally responsible for ensuring the delivery of the outputs as part of good project design and good implementation planning and delivery.	**Output Indicators:** 1. Output indicators have quantity, quality, and time attributes. If time is not stated, the end of project is assumed. 2. Output indicators generally include measures of cost-efficiency. 3. Collection of data for measurement of output indicators is funded by the project. 4. For complex projects, a separate table (perhaps an addendum to Annex 1) may be used to provide a more detailed listing of output indicators. 5. It is better to have only a few meaningful and easily measured output indicators than an abundance of indicators for which data collection is problematic. 6. The output indicators are agreed with the borrower during PCD stage (as to the availability of data, and ease of collection), and a baseline obtained prior to appraisal.	**Project Reports:** 1. Output indicators are generally monitored and/or evaluated via various project reports, supervision mission reports, and evaluation (midterm and final) reports. 2. Sources of data for monitoring and evaluating these indicators typically include administrative and management record keeping systems and summary reports generated by the project.	**(from Outputs to Project Development Objective)** • Assuming that the outputs listed in the far left box are achieved by the end of the project, list any additional assumptions needed to achieve the project objective. • These assumptions may encapsulate conditions, policy changes, or expected behaviors of beneficiary groups or institutions that are necessary for project success. • These assumptions are critical to the achievement of the stated project objective, but are outside the direct control of the project.
Project Components / Subcomponents: 1. A component is a cluster of sub-components or activities that are designed to	**Project Inputs:** (budget for each component) 1. List component inputs in terms of the total cost of each component including	**Project Reports:** 1. Inputs are generally monitored via progress reports and disbursement	**(from Project Components to Project Outputs)** • Assuming that the components and activities listed in the far left box are

produce a single project output. 2. List each project component as a main heading, followed by the major subcomponents, if any, that are funded as a part of it.	contingencies (e.g., US$___) 2. For large or complex projects, the costs for subcomponents may also be shown (indented, to separate them from the component costs).	reports (both quarterly). 2. Inputs are generally evaluated via supervision mission reports (semi-annual) and audit reports (annual).	implemented successfully, list any additional assumptions needed to achieve the stated outputs. • These assumptions are conditions outside the direct control of the project, and are required if the stated project outputs are to be achieved. • The project itself should not be spending money to achieve any of these conditions (since such assumptions are included in the components themselves).

Source: Operational Core Services Department, World Bank. For completed examples of this Annex, visit the M&E Help Desk on the Bank's internal web at http://Lnts012/helpdesk.nsf

Annex 6: Matrix Of Analysis
Nicaragua Emergency Social Investment Fund Impact Evaluation -- 1998

A. Poverty Targeting				
Issues	**General Indicators**	**Methodologies**	**Comments**	**Source of Data**
Poverty levels of SF communities/districts	• % of households in community/district below poverty line and/or consumption levels of extreme poor	requires household income/consumption survey and identification of SF activities by community/district	to compare across countries, need similar definitions of poverty lines	Oversampling national household survey (LSMS) in SF communities -- only for education, health, water and sanitation projects
	• mean consumption level in social fund participant communities vs consumption level in country	requires household income/consumption survey and identification of SF activities by community/district		" "
	• poverty map index (as used by SF)	Maps usually use proxy measures, like a composite poverty index based on mix of variables	Disadvantages are that it arbitrarily chooses indicators and weights, and each country has different index. Advantage is that it often provides more geographical disaggregation than income/consumption surveys - two can be linked to derive predicted consumption	SF uses poverty map based on LSMS93 data using composite poverty index; will update using LSMS98 and Census data to predict consumption at sub-national levels

			levels at dissagreggated levels.	
Poverty levels of SF beneficiaries (household level)	• % of beneficiaries below poverty line or in extreme poverty	income/consumption survey which picks up SF beneficiaries either due to size of SF or by oversampling in SF communities	may vary widely by SF project type	Oversampling national household survey (LSMS) in SF communities
	• mean consumption level of beneficiary households vs. national average for similar households per project type (e.g. with children in primary school, with access to piped water, who use latrines, etc.)	income/consumption survey which picks up SF beneficiaries either by virtue of size of SF or by oversampling in SF communities. Can also oversample in 'match' communities without SF interventions.		Oversampling national household survey (LSMS) in SF communities
Distribution of SF resources	• % of SF projects/resources in bottom quintile of districts	Need consistent ranking methodology across countries		Need to review ranking system and recalibrate

Institutional design features that affect SF targeting performance	• use of poverty map • promotional efforts • direct access by beneficiary groups • share of projects by requesting agency • decentralized offices • target resource allocations • sub-project menu	Develop standard institutional variables that can be used to explain targeting outcomes -- variables easily obtained from SFs		Information available from SF
Other factors affecting targeting performance	• age of SF • 'social capital' of community • distance to SF headquarters • highest education level of beneficiaries • presence of govt/NGO interventions • degree of country income inequality	Also need standard definitions for variables -- variables obtained from SFs, household surveys (with identification of SF beneficiaries), and national surveys		Only indicator which is in doubt is the 'social capital of community'
Comparison of alternatives	• % of SF projects/resources in bottom quintile of districts versus other comparable programs/delivery mechanisms	Compare targeting performance based on geographical location or poverty levels of beneficiaries, depending on survey design, scale of SF and other programs	Difficult to find viable comparators. Need separate information gathering from comparator programs	Planned for Cost-Efficiency Analysis

B. Benefits				
Issues	**General Indicators**	**Data sources/methodologies**	**Comments**	**Case Study: Nicaragua**
Physical capital	• extent to which sub-projects respond to community priorities	Community-level survey, beneficiary assessment, or household survey with oversamples in SF areas		Covered in IDB-financed Beneficiary Assessment and Facilities Survey
	• beneficiary perception of benefit level and improvements to welfare	Household survey or beneficiary assessment in SF communities		Covered in household survey. and IDB-financed Beneficiary Assessment
	• improvement in access to social and economic infrastructure (before and after)	Household survey of SF beneficiaries.	Need to have either baseline or recall questions. Need to develop separate indicators per type of SF project	Some recall questions in household survey. Also possible ex-ante from previous LSMS. Can compare SF beneficiary with characteristics of national population and with match communities.

Physical Capital (cont.)	• improvement in access to social and economic infrastructure versus comparator projects	Household survey of SF and comparator project beneficiaries.	" "	Can compare SF beneficiary with general characteristics of national population as well as match communities.
	• improvement in quality of infrastructure and services (before and after)	Facilities survey and household survey, some coverage from beneficiary assessment	Need to have either baseline or recall questions. Need to develop separate indicators per type of SF project	For education, health, water and sanitation, recall plus historical information from facilities survey and SF ex-ante appraisal
	• improvement in quality of infrastructure and services versus comparator projects	Facilities survey and household survey, some coverage from beneficiary assessment (in SF and comparators)	" "	For education, health, water and sanitation, SF and non-SF facilities through household and facilities surveys

Human capital	• improved educational status - school attendance, years completed, drop-out and retention rates (before and after and versus comparators)	Household survey and information from school		Household survey and information from school for SF and non-SF schools and households
	• improved health status - incidence of disease, infant mortality, malnutrition, increased breastfeeding, etc. (before and after and versus comparators)	Household survey with health module. Anthropometric measures if malnutrition included		Household survey and information from health center for SF and non-SF centers and households
	• improved economic status - increased income, reduced time spent fetching water, lower cost of services, increased employment (before and after and versus comparators)	Household survey		Household survey for SF-beneficiary and non-beneficiary match communities
Social capital	• increased community capacity to address problems (versus comparators)	Household survey, community survey and/or beneficiary assessment		Not addressed
	• increased participation rates in community-initiated changes (versus comparators)	Household survey, community survey and/or beneficiary assessment	Need to develop indicators	Information in household survey on participation.

C. Sustainability of Benefits				
Issues	**General Indicators**	**Data sources/methodologies**	**Comments**	**Case study: Nicaragua**
Sustainability of operations	• conditions under which SF projects are operating after SF intervention (absolute sustainability)	Facilities survey	Can get some additional information from beneficiary assessment	For education and health project surveys, have both SF and non-SF
	• conditions under which SF projects are operating after SF intervention versus comparator projects (relative sustainability)	Facilities survey	" "	" "
Sustainability of maintenance	• maintenance of infrastructure and services over time (absolute)	Facilities survey	" "	" "
	• maintenance of infrastructure and services over time versus comparator projects (relative)	Facilities survey	" "	" "
Sustainability of impact	• quality and quantity of infrastructure and services over time	Facilities survey and household survey	" "	" "
Sustainability of community effects	• tendency of SF communities to submit other proposals (to SF and others) over time	SF database, community survey and/or beneficiary assessment		would need additional work
	• community participation in social and economic infrastructure needs over time	Community survey, household survey and/or beneficiary assessment		include in *next* ben. assessment; impact evaluation

D. Cost Efficiency				
Issues	**General Indicators**	**Data sources/methodologies**	**Comments**	**Case study: Nicaragua**
Cost efficiency of sub-projects	• average cost per new school/health post/water system versus alternative approaches versus comparator projects	SF database and information from government ministries and municipal governments	Costs change over time and comparator projects must be identical	SF and non-SF data from facilities survey. Non-SF cost estimates may not be reliable
	• unit costs: cost per square meter construction, per kilometer of road, etc. versus comparator projects	" "		Can calculate SF averages. Will include in cost-efficiency analysis
	• average cost per beneficiary per SF project type versus comparators	" "		" "
	• average cost of employment generated versus comparators	" "		" "
Cost efficiency of delivery mechanism	• SF institutional costs (investment and operating) as share of SF projects versus comparator projects	" "	Need to develop standard definitions of institutional costs; specify time period	" "
	• average completion time versus comparator projects	" "		" "