Monitoring and Evaluation Technical Notes and Case Studies

TN 1. Major Types of Evaluation

Evaluation is a systematic examination of the relevance, operation and outcomes of programs and policies, compared to a set of explicit or implicit standards, aimed at improving public actions. There are different types of evaluation that address different evaluation questions. These questions can be broadly classified in four categories:

- *Process* questions aimed at understanding what is happening with the program or with a specific component of it, how is it working and whether it is being implemented as originally designed.
- *Outcome* questions seek to assess whether individuals or households' situation (knowledge, behavior, well-being, etc.) have changed. These questions are usually asked in reference to the achievement of the original or revised program goals or to elicit stakeholders views about what the results are, independently from the original intentions.
- Questions of *attribution of outcomes to programs* aim at understanding the extent to which the program is responsible for the observed changes in outcomes. Outcomes may change due to a number of reasons, many of which may be independent of the program. Attribution questions ask whether observed changes are caused by the program or whether they would have occurred anyway.
- *Questions on reasons* aim to explore the reasons behind the observed process and outcomes; they ask how and why results were what they were

These questions can be roughly matched by three major types of evaluation: process evaluation, outcome evaluation, and theory-based evaluation. Each type of evaluation in turn has a menu of possible evaluation designs and data collection methods. Evaluation designs are bundles of techniques that can be used in different combinations to answer different evaluation questions. Evaluation designs specify the units of analysis (e.g. households, individuals, facilities, communities, etc.) and how they are going to be selected (opportunistically or using systematic sampling strategies); which kind of comparisons will be made (e.g. no comparison, comparison across time or space, between groups, etc.); and the timing of the data collection (e.g. before and after the program, immediately after the program only, during program implementation, etc.).

Process evaluations assess how effectively a public action is being implemented; they focus on aspects such as who is participating, what activities have been offered, what actions have been taken, what are staff practices and client responses. A process evaluation can be conducted for several reasons including when problems such as delays, cost overruns, or beneficiary dissatisfaction have been detected by the monitoring system or on a regular basis as an early warning system. Process evaluations tend to rely on less formal evaluation designs and modes of inquiry such as self-evaluation and expert judgement.

Outcome evaluations assess what happened to individuals (or other unit of analysis) after the policy or program implementation; they focus on intervention results such as whether people are healthier, better educated, and less vulnerable to adverse shocks. Evaluation designs for outcome evaluations vary along a continuum of levels of complexity. At the one end of the spectrum are outcome evaluations that simply assess whether program participants experienced any changes in key welfare indicators – these are basically monitoring exercises. Evaluation designs and data collection and analysis methods on this side of the spectrum tend to be relatively simple and quick

to yield results, but leave room for differing interpretations of how much change has actually occurred and how much of it can be attributed to a particular intervention. Evaluation designs generally look only at the group of program participants; there is no comparison with people or communities that did not participate in the program or any efforts to isolate program or policy effects from other events occurring simultaneously. The evaluation can look at outcomes either only after the intervention has been in operation for a while or is completed, or before and after the intervention. Data collection and analysis methods can be quantitative such as rapid service delivery surveys; qualitative such as key informant interviews or focus groups; or participatory such as rapid appraisal methods.

At the other end of the spectrum are evaluations that address attribution questions using special – often complex -– techniques to disentangle the net gains from interventions (see **Technical Notes 2 and 3**). These evaluations are usually referred as impact evaluations. Impact evaluations assess the extent to which public actions have produced their intended effects and the extent to which changes in individuals' well-being can be attributed to a particular program or policy. They estimate the magnitude of the effects of a program/policy and assign causation. Such a causal analysis is essential for understanding the relative role of alternative program interventions in reducing poverty, and thus for designing appropriate poverty reduction strategies.

Theory-based evaluations examine the link between inputs, activities and outcomes and aim at determining whether a breakdown has occurred, where, why and how. They present the explicit or implicit theory about how and why a public action would work as a series of microsteps, and analyze them sequentially to track the unfolding of assumptions. By following the sequence of steps, this type of evaluation can determine if and where the process from program inputs to outcomes broke down.

TN 2. Impact Evaluation Designs

Experimental or randomized designs involve gathering a set of individuals (or other unit of analysis) equally eligible and willing to participate in the program and dividing them into two groups: those who receive the intervention (treatment group) and those from whom the intervention is withheld (control group). For example, in some social funds, economically feasible projects submitted by communities are randomly selected to receive funding during the first phase of the project (treatment group), while the rest are scheduled to receive funding at a later stage and can thus be used as control group. Since program participants are selected randomly, any difference with non–program participants is due to chance. For this reason, experimental designs are usually regarded as the most reliable method and the one yielding the easiest-to-interpret results. In practice, however, this type of evaluation design can be difficult to implement, not least because it is difficult to withhold benefits from equally eligible individuals (see **Case Studies 5 and 11**).

Quasi-experimental design is another option. When randomization is not feasible, a comparison group can be constructed. The two methods for constructing a comparison group are matching and reflexive comparisons. Matching consists of selecting non–program participants comparable in essential characteristics to participants, on the basis of either a few characteristics or a number of them, using statistical techniques. For example, the evaluation of Trabajar, a public works program in Argentina, constructed a comparison group by matching program participants to nonparticipants on the basis of several socioeconomic characteristics, including schooling, gender, housing, subjective perceptions of welfare, and membership in political parties (see **Case Study 4**). Evaluations using matching methods are often easier and cheaper to implement than experimental designs, but the reliability of results is lower and the interpretation of results is more difficult.

Another type of quasi-experimental design is called reflexive comparison. In a reflexive comparison, the counterfactual is constructed on the basis of the situation of program participants before the program. Thus, program participants are compared to themselves before and after the intervention and function as both treatment and comparison group. This type of design is particularly useful in evaluations of full-coverage interventions such as nationwide policies and programs in which the entire population participates and there is no scope for a control group (see **Case Study 8**). There is, however, a major drawback with this method: the situation of program participants before and after the intervention may change owing to myriad reasons independent of the program. For example, participants in a training program may have improved employment prospects after the program. While this improvement may be due to the program, it may also be due to the fact that the economy is recovering from a past crisis and employment is growing again. Unless they are carefully done, reflexive comparisons may not be able to distinguish between the program and other external effects, thus compromising the reliability of results.

Nonexperimental designs can be used when it is not possible to select a control group or a comparison group. Program participants can be compared to nonparticipants using statistical methods to account for differences between the two groups. Using regression analysis, it is possible to "control" for the age, income, gender, and other characteristics of the participants. As with quasi-experimental methods, this evaluation design is relatively cheap and easy to implement, but the interpretation of results is not straightforward and results may be less reliable.

TN 3. Impact Evaluation Methods for Policies and Full-Coverage Programs

Computable general equilibrium models (CGEs) attempt to contrast outcomes in the observed and counterfactual situations through computer simulations. These models seek to trace the

operation of the real economy and are generally based on detailed social accounting matrices (SAMs) collected from data on national accounts, household expenditure surveys, and other survey data. CGE models do produce outcomes for the counterfactual, though the strength of the model is entirely dependent on the validity of the assumptions. This can be problematic, as databases are often incomplete and many of the parameters have not been estimated by formal econometric methods. CGE models are also very time consuming, cumbersome, and expensive to generate.

With and without comparisons compare the behavior in key variables in a sample of program countries or regions to their behavior in non–program areas (a comparison group). Thus, this method uses the experiences of the non–program areas as a proxy for what would otherwise have happened in the program countries. An important limitation of this approach is that it assumes that only the adoption of a particular policy or program distinguishes program countries or regions from non–program areas and that the external environment affects both groups equally.

Statistical controls consist of regressions that control for the differences in initial conditions and policies undertaken in program and non–program countries or regions. The approach identifies the differences between program and non–program areas in the preprogram period and then controls these differences statistically to identify the isolated impacts of the programs in the postreform performance.

Source: Adapted from Baker 2000.

TN 4. Types of Data Sources for Impact Evaluation

Longitudinal or panel data sets include information on the same individuals (or other unit of analysis) at least at two different points in time, one before the intervention (the baseline) and another afterwards. Panel data sets are highly valued for program evaluation, but they can be expensive and require substantial institutional capacity (see **Case Study 8**, and the chapter on **Poverty Data and Measurement**).

Cross-section data contain information from program participants and nonparticipants at only one point in time, after the intervention. Evaluations using cross-section data usually cost less than studies using information from more than one point in time, but the results tend to be less reliable, except for experimental designs (see **Case Study 4**).
Before-and-after, or baseline with follow-up, data consist of information at two points in time: before and after the intervention. These surveys may or may not include data on non–program participants. If the evaluation is based on a simple before-and-after comparison of program participants (reflexive comparison), the results should be interpreted with caution (see **Case Study 10**).

Time-series data gather information on key outcome measurements at periodic intervals both before and after the program. They allow the examination of changes in trend pre- and postprogram. However, many data points before and after the program are required for rigorous analysis. Time series are mainly used to evaluate policies and programs with full or national coverage.

Case Studies

Case studies 1 and 2 provide examples of national poverty monitoring systems, whereas case study 3 presents an example of the use of citizen feedback surveys as a tool for civil society participation in assessing public sector performance. Case studies 4 to 12 (adapted from Baker 2000) exemplify impact evaluations of projects and programs across different sectors. They illustrate a wide range of approaches in evaluation design, use of data, policy relevance of results, and associated impact on evaluation capacity building (see table 7). Each study includes a discussion on the relative strengths and weaknesses of each evaluation.

Table 7. Summary of Impact and Evaluation Case Studies

					Econometric Approach					
Program/project	*Country*	*Database type*	*Unit of analysis*	*Outcome measures*	*Random-ization*	*Matching*	*Reflexive comparisons*	*Instrumental variables*	*Qualitative evaluation*	*Strengths*
Education										
School autonomy reform	Nicaragua	Panel survey and qualitative assessments	Students, parents, teachers, directors	Test scores, degree of local decisionmaking	No	Yes	Yes	No	Yes	Qualitative-Quantitative Mix
Dropout intervention	Philippines	Baseline and post intervention survey	Students, classrooms, teachers	Test scores and dropout status	Yes	No	Yes	Yes	No	Cost/benefit analysis; capacity building
Labor Programs										
Trabajar program	Argentina	Household survey, census, administrative records, social assessments	Workers, households	Income, targeting, costs	No	Yes	No	Yes	Yes	Judicious use of existing data sources, innovative analytic techniques
Active labor programs	Czech Republic	Retrospective mail surveys.	Workers	Earnings and employment outcomes	No	Yes	No	No	No	Matching technique
Finance										
Credit with education	Ghana	Baseline and post intervention survey	Mother/child pairs	Income, health and empowerment	Yes	Yes	Yes	No	Yes	Use of qualitative and quantitative information
Health										
Health financing	Niger	Baseline and post intervention survey plus administrative records	Households and health centers	Cost recovery and access	No	Yes (on districts)	Yes	No	No	Use of administrative data
Infrastructure										
Social investment fund	Bolivia	Baseline and follow-up surveys	Households, projects	Education and health indicators	Yes	Yes	Yes	Yes	No	Range of evaluation methodologies applied
Rural roads	Vietnam	Baseline and follow-up surveys	Households, communities	Welfare indicators at household and commune levels	No	Yes	Yes	Yes	No	Measures welfare outcomes
Agriculture										
National extension project	Kenya	Panel data, beneficiary assessments	Households, farms	Farm productivity and efficiency	No	No	Yes	No	No	Policy-relevance of results

Source: Adapted from Baker 2000.

CS 1. Monitoring the Progress of the Poverty Eradication Action Plan in Uganda[1]

1. Introduction

In 1995 the Government of Uganda embarked in the formulation of the Poverty Eradication Action Plan (PEAP) to ensure that poverty reduction was the major focus of its overall growth and development strategy. This plan was formulated through a consultative process involving representatives from the government and civil society as well as donor organizations. The overarching goal of the PEAP is to eradicate mass poverty – reducing the percentage of the population living in absolute poverty from 56 percent (1992) to 10 percent, and cutting the percentage of people living in relative poverty from more than 85 percent to 30 percent by 2017.

Additional goals were agreed on in four areas – macro-economics, governance, income generation and human development – and expanded into a set of strategic objectives (see Box 1). Primary health care, primary education, agricultural extension and rural feeder roads were identified as initial priority poverty areas for resource allocation. Goal setting and the choice of strategic objectives and priority areas have been dynamic processes, frequently revised in light of new information – such as the Uganda Participatory Poverty Assessment (UPPA) conducted in 1998 and 1999 – and feedback from the poverty monitoring system.

Box 1: Poverty Eradication Action Plan Goals and Strategic Objectives

Overarching goal: To reduce the percentage of the population living in absolute poverty from 66 percent to 10 percent and cutting the percentage of people living in relative poverty from more than 85 percent to 30 percent by 2017

Goal #1. Implementation of macro-economic policies that provide an enabling environment for poverty reduction

- Maintain a stable exchange rate and one that makes the export sector competitive
- Maintain low levels of inflation that facilitate savings mobilization and long-term planning
- Promote private sector investment in rural areas
- Reduce anti-export bias of trade policy to improve prospects for exports
- Promote broad based economic growth
- Reduce external indebtedness to sustainable levels
- Reduce poverty disparities among districts
- Improve women's economic and political empowerment
- Broaden tax base
- Re-focus public expenditure to be directly linked to poverty eradication

Goal #2: Creation of an institutional framework that promotes poverty reduction through broad participation, transparency and accountability

- Enhance the effective and efficient delivery of public services while fostering transparency and accountability
- Promote the growth of the private sector by enhancing local and foreign investments
- Strengthen the machinery for keeping law and order and administering justice, while improving poor people's access to legal services
- Enhance the observance of human rights and freedom and democratic governance

- Promote community participation in the planning and delivery of services

Goal #3: Expansion of the income opportunities of the poor

- Provide an efficient road network
- Transform/modernize agricultural production
- Ensure security of land tenure; adequate accessibility to land and its efficient use, while preserving the environment.
- Support development of rural markets: infrastructure, market information, and standards
- Provide financial services to the poor through promotion of the growth of micro financial institutions and rural village banks
- Ensure security of tenure to all, to enhance effective and sustainable land use
- Enhance labor productivity giving priority to employment of women, reduction of the exploitation of child labor, and safeguard the rights of employees
- Create an enabling environment for the development of micro and small scale enterprises

Goal #4: Improvement of the quality of life and the human capital of the poor

- Meet the constitutional provision of basic health care to all, improving the delivery of health services to the entire population on a cost-effective basis
- Provide safe drinking water to the entire population within easy reach, while improving cost-effectiveness of water provision
- Achieve universal primary education and improve the quality of education
- Promote access to basic education for vulnerable children (e.g. the homeless and street children)
- Promote the acquisition, use and retention of functional literacy by all the people of Uganda

2. Poverty Monitoring System

Progress in achieving the goals is being assessed through continuous poverty monitoring. This started as an ad-hoc activity and has evolved gradually towards a decentralized, participatory monitoring system with a clearer delineation of roles and responsibilities, including mechanisms for collaboration across institutions.

The system consists of three core elements:

- The Uganda Bureau of Statistics (UBoS), which collects, analyzes and publishes data from household surveys.
- The Statistics Departments in line ministries, which collect and analyze sectoral data from management information systems.
- The Poverty Monitoring Unit, whose main function is to link data producers and policy-makers. It collects poverty data from different sources including UBoS, line ministries, and other organizations and institutions outside the government; analyzes the data; disseminates results, and discusses poverty trends and outlooks with government representatives and bodies. In the future, the Unit will expand to include policy analysis for poverty reduction in its mandate. The Unit sits in the Ministry of Finance, Planning and Economic Development, which is key for influencing policy.

In addition, other institutions such as non-governmental organizations, academic institutions, research centers and donors play an important, but not yet systematic, role in collecting and analyzing additional data. Policy-makers are also a key part of the system as the main users of monitoring results (primarily at the central level, although it has been recognized that locally collected statistics must also be used in local decision-making).

[1] This case study was prepared by Margaret Kakande, Poverty Analyst, Poverty Monitoring Unit, Ministry of Finance, Planning and Economic Development, Government of Uganda; Kimberley McClean, Head International Projects, Aus Health International; and the chapter authors.

The system is undergoing a major revision aimed at:

- Increasing participation – i.e. promoting higher involvement in monitoring activities at the local level and collaboration between the Uganda Bureau of Statistics (UBoS), the Poverty Monitoring Unit, non-government organizations and line ministries in collection, analysis and dissemination of data. Linkages between the districts and central bodies collating statistics are also being revised.
- Developing capacity – particularly for monitoring at local (district) levels, and in data analysis and dissemination at central levels in order to decrease the lag time between data collection and analysis/dissemination.
- Defining institutional roles – i.e. setting clearly defined roles and responsibilities, including mechanisms of collaboration.
- Harmonizing progress reporting – i.e. defining a common format for sectoral and poverty programs progress reporting

One of the options under consideration to address some of these issues is the establishment of a field organization for the UBoS. The field organization would be responsible for controlling the flow of information to and from headquarters; backstopping the development of district statistics; recruitment, training and supervision of field staff; scheduling of field work; actual data collection and data entry and carrying out all other functions associated with field work. Six statistical zones would be established. Each zone would have a zonal office with a small number of permanent staff (zonal supervisor, statistical assistant and data entry operator) plus field supervisors and enumerators that would be recruited on a temporary basis.

Indicators

The selection of indicators has been an iterative process. Originally, indicators were selected based on the work of thematic groups to monitor progress in a number of areas: income poverty; health status; education; environment, infrastructure, governance, employment, and access to information, markets and credit. The first list of indicators was perceived as too long, incomplete in some areas and lacking focus against the goals and strategic objectives. As part of the Medium Term Expenditure Framework (MTEF), the Poverty Working Group – composed of government officials and representatives from civil society and donor organizations – refined the list of indicators (see Box 2). This list will be further adjusted to ensure continued consistency with the revised PEAP. Nearly all indicators are currently monitored nationally; a subset is monitored at the district and/or regional level. Education information is the only data that are disaggregated by sex. This is a major limitation for a complete poverty analysis and is expected to be addressed during the current review process.

Although some progress has been made in aligning the indicators with the goals and strategic objectives, there are still areas for improvement. Several indicators are defined in terms of number of cases. Actual numbers are important, but in many cases percentages and ratios can make indicators more useful. For example, the proportion of health units with essential drugs is a more informative indicator than just the number of units. Another problem with some indicators is that they are not unambiguous measures of progress, i.e. it is not possible to determine whether the situation has improved or not based on that indicator. For example, an increase in household expenditures in education is not an unequivocal indication of improvement. Households maybe spending more on education because they consume more or because they have to pay more and maybe consume less. Finally, the list does not distinguish between final and intermediate indicators, which would be useful when judging overall progress.

TABLE 1: Revised list of monitoring indicators

Indicators	Intended level of disaggregation
INCOME POVERTY	
Proportion of population below the poverty line	National, regional, district
Number of people in absolute poverty	National, regional
Household percentage share of food expenditure	National, regional
Proportion of population living under thatched houses	National, regional
Dependency ratio	National, regional, district
Gini coefficient	National, rural/urban
Consumption per capita of poorest 20%	National, regional, district
Per capita GDP	National
Savings/GDP ratio	National
Revenue per capita per district	District
Security and vulnerability	
Proportion of households affected by theft or civil disturbance	National, regional
Number of people internally displaced	National, regional
Number of civilian deaths due to insurgency	National, regional
Number of criminal cases reported	National, regional
Proportion of households experiencing major income shocks last year	National, regional
Refugee and displaced as proportion of district population	District
Proportion of households under economic distress selling assets	National
Road Network	
Road length opened	National
Road length up-graded	National
Proportion of districts with more than 50% of roads in poor condition	National, district
Proportion of area not serviced by roads	National, district
Land	
Incidence of poverty by land ownership and tenure	National, district
Agriculture	
Adoption rate of modern farming methods	National, district
Yield rates	National, district
Percentage of farmers growing food security crops	National, district
Markets	
Availability of markets by type	National, district
Accessibility of markets	National, district
Volume of goods and services handled at a given market	National, district
Proportion of households where the sale price of the main agricultural product is less than 50% of the urban market price	National, district
Labor productivity and employment	
Unemployment rate	National, district
Vocational training enrollment	National, district
Average hours worked per day	National, district
Rural credit	
Growth in micro-finance portfolio	National, district
Proportion of population accessing micro-credit	National, district
Growth in savings	National, district
Credit management (effective use)	National, district
Availability of micro-finance services	National, urban/rural

Indicators	Intended level of disaggregation
QUALITY OF LIFE	
Health Indicators	
Incidence of disease	National, district
Immunization coverage	National, district
Proportion of population with 5km to the nearest health unit	National, district
Per capita household expenditure on health	National, district
Number of health units with essential drugs	National, district
Number of districts with more than 1,000 people per trained health personal	National, district
Antenatal care coverage	National, district
Water and Sanitation	
Proportion of population within ½ km to safe water by region	National, district rural/urban
Proportion of population with good sanitary latrines	National, rural/urban
Safe waste disposal	National
Education indicators	
Net primary enrollment ratio	National, district, gender
Proportion of primary school pupils completing more than 4 years of education	National, district, gender
Pupil/trained teacher ratio	National, district, gender
Distance to schools	National, district, gender
Pupil/classroom ratio	National, district, gender
Pupil/textbook ratio	National, district, gender
Per capita household expenditure on education	National, district, gender
ENVIRONMENT	All National
Level of compliance to environmental standards	
Corrective actions by NEMA	
Proportion of the population practicing sustainable land-use methods	
Budgetary allocations to environmental programs by local governments	
Proportion of gazetted land in districts	
GOVERNANCE AND ACCOUNTABILITY	
Level of awareness among the population on rights/entitlements	National
Proportion of reported cases cleared	National
Number of people on remand beyond the specified period by law	National
Number of backlog court cases	National
Corruption cases raised at different levels	National
Successful programs in poverty eradication	National, district
Number of corruption/embezzlement and abuse of office cases resulting into conviction	National

Data Collection

Main data sources for monitoring include household surveys, management information systems and qualitative studies.

Household surveys are centrally planned and implemented by UBoS with limited consultation or participation at the district level. The role of districts is under review, with the objective of building local capacity and promoting rapid access to district-specific information that can be used by districts for planning, implementing and monitoring of their programs and policies. UBoS and the Ministry of Local Government are working on a system involving the District Planning Units in data collection to ensure that relevant statistics and qualitative information are utilized to monitor performance at the district level. Such a system would complement the household data collection system that is managed centrally. Household surveys for poverty monitoring include:

- *Integrated Household Surveys (IHS),* which collect data on household characteristics, housing characteristics, household income and expenditures, assets, loans and savings, agricultural production, health and nutritional status of children. The IHS conducted in 1992 provided baseline information on 10,000 households throughout the country. The survey questionnaire

was revised based on insights from the Uganda Participatory Poverty Assessment, and now includes questions on topics such as household security. The revised survey – the Uganda National Household Survey – was conducted in 1999/2000.

- *Monitoring Surveys* that collect information similar to the IHS using a smaller sample of 5,000 households and a smaller questionnaire (which does include a consumption module). They have been conducted annually from 1992/93 to 1997.
- *Demographic and Health Surveys* collect information on maternal and child health, immunization, health care access, major disease incidence, etc. Baseline data were provided by the 1995 DHS; a follow up survey is expected in 2001.

Also, the *Population Census 2002* will provide updated information on the demographic structure of the population – age, marital status, ethnicity, religion, household size, dependency ratios, etc.

Other surveys such as the Public Expenditure Tracking Survey (see Public Spending Chapter) and the National Service Delivery Survey have provided useful information but have not yet become part of the regular monitoring system. The National Service Delivery Survey collects information on usage of and satisfaction with public services. These surveys were piloted in 1996 and conducted nation-wide in 1999 (currently by the Ministry of Public Services but in the future by the Bureau of Statistics).[2]

Management Information Systems (MIS) collect sectoral information on outputs, access to services, and to a limited extent, on quality of services. For example, in the health sector, the MIS gathers information on the number of health facilities by type, public or private management and bed capacity; facilities offering essential services; staffing; and major causes of morbidity. For education, the Ministry of Education and Sports conducts an *annual education census* using district-level information on enrollment of pupils, teachers, teaching/learning materials, facilities and finances.

A number of problems with the MIS data have been identified in Uganda. First, information is incomplete. By 1996, for example, the education census had a 60 percent response rate from government-assisted institutions, and 30 percent from private schools. Second, data are not reliable. In the education sector, reliance on head teachers to provide school data is problematic – student numbers are often inflated in order to obtain higher grants. Random checks have been implemented and reveal enrollment over-reporting. In the health sector, diagnostic tools, staff capacity and communication infrastructure are limited in many areas – especially remote rural areas – so that gross under-reporting of disease incidence occurs. Furthermore, there is an issue of timeliness; the arrival of data from districts is slow, and data analysis, compilation and reporting at the center is delayed so data are not used for service provision and planning. As a result, for example, the Education Statistics Abstracts are usually produced 1.5 year after data collection.

One of the reasons identified for the low performance of Management Information Systems is their high level of centralization. Districts are required to collect information without being involved as stakeholders in the monitoring process. Hence, they have few incentives to ensure the timely collection of reliable data. Efforts to correct this situation comprise activities at the district level and at the sector level with central line ministry. District level activities include the implementation of the **District Resource Information System (DRIS)**. DRIS is the second phase of an earlier attempt to collect data on social services and relevant infrastructure from all districts (the District Resource Endowment Profile Study, or DREPS). It establishes a direct link between districts and UBoS and focuses on a larger number of variables including administration, service delivery and infrastructure. Unfortunately, DRIS does not include agriculture data. This poses a serious

[2] For more information on the National Service Delivery Survey, see the World bank Research Department web site on public service delivery:
http://www.worldbank.org/research/projects/publicspending/tools/tools.htm#Quantitative Service Delivery.

limitation for monitoring poverty reduction as the majority of the poor are engaged in the agricultural sector.

The Uganda Participatory Poverty Assessment Project (UPPAP) is the main source of qualitative data for PEAP monitoring. It is a three-year project aimed at incorporating the perceptions of poor people into the local and national dialogue for poverty reduction and providing a deeper understanding of trends emerging from quantitative data. The next participatory poverty assessment is planned for 2001. The UPPAP is a partnership between the Government, donors, the nine district authorities in which the project operates, and Oxfam, as the implementing agency. Within the government, the UPPAP is situated into the Ministry of Finance, Planning and Economic Development under the Poverty Monitoring and Analysis Unit.

Agriculture and governance are two areas in which additional work is needed. As mentioned earlier, despite the large proportion of poor people engaged in agricultural activities, agriculture data are not readily available. Likewise, there is little usable information on governance issues except from the **National Integrity Survey** and **Reports from Human Rights Commission**. There is a wealth of data collected by the Inspectorate General of Governance, but as they are understaffed no summary statistics are available.

Data Analysis, Dissemination and Feedback

Data analysis is conducted mostly at the central level by UBoS. At the district level, data analysis is limited and there is no real reporting (but there are plans to strengthen the field structure of the UBoS). A few districts have started their own monitoring systems under the Local Government Development Project, whose main objective is to strengthen participatory planning and the development of budgeting and monitoring systems at the district level. The project has recently been extended from the original 5-9 districts to a further 30.

A Poverty Status Report is produced every two years by the Poverty Monitoring Unit to assess progress and challenges in the implementation of the PEAP. It provides an overview of progress towards the PEAP goals as well as the status of poverty eradication actions, including budget allocations. This information sets the basis for identifying gaps, key challenges and priority areas.

Reports are disseminated at the national and district level and are used in the revision of the PEAP, the MTEF and sector reviews. The Poverty Monitoring Unit and the Poverty Working Group (PWG) act as a link between the monitoring results and the policy making and budget allocation processes. Specifically, the PWG integrated by government officials, representatives of NGOs and academia, ensures that the data collected from the poverty monitoring system are taken into consideration and acted on by the relevant sector working group in the MTEF and budget processes. It also makes recommendations on the overall budget allocation of resources for poverty reduction as well as on other budget policies that have an impact on the poor.

Although some efforts have been made to establish a link between poverty monitoring activities and policy making and budget allocation, the performance of public expenditures is still mostly measured in terms of inputs and activities rather than contributions to poverty reduction. Progress towards goals still plays a limited role on the sectoral budget allocation process. An incentive system linking resource allocation and performance assessment to contributions to PEAP outcomes needs to be developed.

3. Statistical Capacity Building

The activities of the poverty monitoring system are supported by a major program to upgrade Uganda's statistical systems. The main goal of the program is to build national capacity to collect, process, store and disseminate statistical information for poverty monitoring and evaluation at both the national and district levels. The program focuses on strengthening the capacity of UBoS to deliver a core statistical program that allows a regular and timely monitoring of national

development goals. It establishes a new information technology infrastructure for an *Integrated Information Management System.* This system is designed to ensure that all data collected directly by UBoS or received by UBoS as secondary data from other sources are centrally stored in a common format that will facilitate open access to the data by users, whether in hard copy or electronic form. The Central Depository of Data holds all the data in a cleaned format ready for use, thus guaranteeing that all tables and analysis are based on the same data source. The system also incorporates the concept of a centralized store of macro data or output tables which is then used as input source of reports, newsletters or electronic dissemination.

Another area of special interest is upgrading UBoS's household survey capabilities. The main activities include a three-year Integrated Household Survey Strategy and program and the establishment of a core field force of several mobile teams. These teams will be used both to conduct UBoS surveys as well as to serve as a pool of technical support for districts planning their own surveys. UBoS will conduct a pilot study on a simple indicators monitoring surveys that could be carried out by district governments to meet their information needs.

The program will also support the repeated administration of an annual National Service Delivery Survey (NSDS). This survey uses a small questionnaire and a large sample (approximately 20,000 households) so as to be able to disaggregate results at the district level. It incorporates a number of features of the Core Welfare Indicators Questionnaire including the use of Optical Mark Recognition to speed up the data entry process. As mentioned earlier, the survey will be progressively mainstreamed and taken over by UBoS. Future rounds of the survey will be supplemented with focus groups interviews.

4. Sources

Hauge, Arild. 2001. "Strengthening Capacity for Monitoring and Evaluation in Uganda: A Results Based Management Perspective." ECD Working Paper Series, No. 8. World Bank, Operations Evaluations Department, Washington, D.C.

Republic of Uganda. 1997*. Poverty Eradication Action Plan: A National Challenge for Uganda*. Ministry of Planning and Economic Development, Kampala

Republic of Uganda. 1999. *Uganda Poverty Status Report*, 1999. Ministry of Finance, Planning and Economic Development, Kampala

--------- 1999. *Five Year Strategy for Poverty Monitoring and Policy Analysis.* Planning and Poverty Eradication Section, Ministry of Finance, Planning and Economic Development, Kampala

The World Bank. 1999. "Uganda's Integrated Information Management System: A New Approach in Statistical Capacity-Building." Findings, No. 142, Washington, D.C.

CS 2. Proposed Plan to Monitor the Poverty Reduction Strategy in Tanzania

1. Introduction

The Poverty Reduction Strategy in Tanzania builds upon earlier strategies to address poverty and enhance human development. It consolidates previous medium and long-term strategies such as Vision 2025, the 1997 National Poverty Eradication Strategy (NPES) and the Tanzania Assistance Strategy and lays out a plan focused on three broad goals:

- reducing income poverty;
- improving the quality of life and social well-being; and

- achieve and sustain a conducive development environment

The preparation of the PRS has been characterized by broad-based participation of stakeholders. Throughout the process, the views of grassroots stakeholders including local governments, local communities and civil society were gathered through zonal workshops. The draft targets, priorities, and actions were also discussed at a national workshop, which included central and regional government officials, private sector organizations, the donor community and the media.

2. Monitoring the Poverty Reduction Strategy

The Poverty Reduction Strategy Paper presented a tentative plan to monitor and evaluate the strategy. This plan has been refined since the launch of the PRS and will continue to evolve as new lessons emerge during implementation. This case study highlights three aspects of the Tanzania experience: selection of indicators, data sources and the planned institutional framework for monitoring and evaluating the PRS. Other activities, such as participatory studies, reporting of results and advocacy work, are not highlighted here.

Selection of Indicators

The M&E system includes a set of final and intermediate indicators. Final indicators were selected from a wider list of poverty and welfare indicators resulting from a consultative process, and will be used to monitor progress toward the main goals of the strategy. Intermediate indicators will be used to monitor the implementation of the strategy in terms of resources allocated and the goods and services generated through key policy actions. In recognition of the difficulty of measuring some final indicators at frequent intervals, the monitoring system also includes a set of proxy indicators that can be monitored on an annual basis. For example, one of the objectives of the PRS is to reduce income poverty. Thus, monitoring the proportion of the population living below the poverty line at regular intervals is important. However, in the case of Tanzania, as in many other countries, collecting income or expenditure data at frequent intervals was not feasible, so it was decided to include indicators of ownership of household assets and construction materials of dwelling units - that can be monitored annually - as proxy indicators for income poverty.

As shown in Table 1, the proposed indicators fall in four areas broadly in line with the objectives of the PRS: income poverty, quality of life and social well-being (health, education, vulnerability and social well-being), macroeconomic stability and governance. The adequacy of indicators in terms of relevance, clarity, reliability, timely availability and balanced mix between final and intermediate indicators varies greatly across areas.

The PRS chooses an appropriate country-specific final indicator for income poverty, the incidence of income poverty measured on the basis of the national poverty line, and a more ambitious target than under the IDGs: halving the incidence by 2010 instead of 2015. The incidence of poverty will be disaggregated by rural and urban areas; while still largely a rural phenomenon, income poverty is increasingly becoming an urban problem. As mentioned, since income poverty is not measured every year, proxy indicators have been identified, but to ensure timeliness it will be necessary to better define future plans for data collection. The intermediate indicators chosen are relevant in the Tanzanian context, and should also be available on an annual basis for PRS review. There is a good mix of intermediate and final indicators.

Health, survival and nutrition indicators capture the overarching goal of raising life expectancy to 52 years by 2010. However, some intermediate indicators could be defined better to be more informative. For example, implementation of the malaria control program and implementation of the integrated management of childhood illness program are not well defined, unless they refer to specific lists of indicators contained in other documents, such as sector or program monitoring plans. If they do not refer to indicators specified elsewhere, they should be defined more clearly; for example, the indicator on the implementation of the malaria control program could be rephrased as the proportion of primary and secondary health care facilities with a regular supply of

1st and 2nd line antimalarial drugs. Likewise, the percentage of primary and secondary health care facilities with personnel trained in Integrated Management of Childhood Illness (IMCI) is a better indicator than whether or not the IMCI Program has been implemented. Timeliness may be an issue with some of the final and a few intermediate indicators (for example breast feeding practices), since the main data source for these indicators is the Demographic and Health Survey (DHS) that is expected to be conducted during the implementation period of the PRS, but the exact timing of which is as yet unknown.

Education indicators and targets are relevant to the goal of eradicating illiteracy by 2010 – raise gross primary enrolment to 85%; increase the transition rate from primary to secondary school from 15 to 21%; reduce the dropout rate in primary school from 6.6 to 3%; raise net primary school enrolment from 57 to 70%. Most of the final indicators are clearly defined, except for gender equity which could be measured with respect to gross enrolment rates, net enrolment rates, illiteracy rates or some other indicator. The list of intermediate indicators appears incomplete: indicators of outputs from public expenditures, such as pupil-teacher ratio, textbook availability, percentage of classrooms rehabilitated, average travel time to school could supply useful information that help understand trends in final indicators. Education indicators are likely to be available at frequent intervals since they are obtained from routine data collection systems of the Ministry of Education. Enrolment data will be validated with information from the 2002 Census.

Vulnerability indicators are less well-developed. They do reflect the policy actions that will be implemented in this area but monitor activities rather than results – no "final indicator" is really included. More specific intermediate indicators would need to be developed. For example, the percentage of farmers in drought-prone areas switching to drought-resistant crops may be a better indicator than whether or not the production of drought-resistant crops has been promoted. Likewise, a measure of use of the database on vulnerable groups could be a more useful indicator than whether or not such data base has been developed.

Social well-being indicators also reflect the difficulty of specifying measurable indicators. A multiplicity of issues is addressed under this heading. These indicators try to capture progress in the devolution of responsibilities for key services to local authorities; access to justice, efficiency and transparency of the administrative system and the level of participation of all stakeholders in the PRS process, but will give only a very partial picture of progress on these themes. This is an area where goals, indicators and targets would need to be developed further.

Macroeconomic and governance indicators aim to measure the extent to which an environment conducive to development has been achieved. Specifically on the macroeconomic side the PRS aims to attain an inflation rate broadly in line with the anticipated inflation of Tanzania's main trading partners. This goal complements the objective of reaching a 6 percent GDP annual growth over the next three years that would set the basis to achieve the medium and long term poverty reduction goal. The proposed intermediate and final macroeconomic indicators will provide relevant information for monitoring the progress on the stability goal at frequent intervals.

On the governance side, the main goals are to improve the performance of the public sector including the delivery of public services; minimize resource leakage; and promote accountability. The proposed indicators to monitor these goals are in general not well-developed. Unlike in other areas, not all the items listed under "final indicators" are indicators; for example, "a governance system that is efficiently and effectively decentralized" is an objective, not an indicator of decentralization. Several other items are not measurable indicators; for example, "spread and magnitude of corruption" does not identify how corruption would be measured; indicators relevant to Tanzania should be identified.

Table 1 Proposed indicators for monitoring the PRSP in Tanzania

Objectives	Final indicators	Intermediate indicators
1.Reducing income poverty	- Poverty incidence **Proxy indicators** - Ownership of household assets - Type of construction materials of dwelling units (floors, walls, and roofing).	- Real GDP growth - Investment (physical and human) - Investment productivity - Growth in value-added of agriculture - Development of Private Sector Strategy - Seasonal production of key food and cash crops - Kilometers of rehabilitated rural roads - Actual and budgetary allocation for rural roads - Actual and budgetary allocation for agricultural extension
2.Improving quality of life and social well being		
A. Health, survival and Nutrition	- Infant and under-five mortality rates - Percentage of children under 2 years immunized against measles and DPT - Seropositive rate in pregnant women - Maternal mortality - Life expectancy - Malaria-related fatality rate for children under five - Burden of disease/morbidity - Proportion of households with access to safe drinking water - Stunting prevalence - Wasting prevalence	- Proportion of districts with active HIV/AIDS awareness campaigns - Percentage of births attended by trained personnel - Child feeding practices - Implementation of malaria control program - Implementation of Integrated Management of Childhood Illness program - Actual and budgetary allocation for primary health care - Actual and budgetary allocation for HIV/AIDS - Actual and budgetary allocation for water and sanitation
B. Education	- Illiteracy rate - Gender equality in primary and secondary education - Proportion of school age children successfully completing primary education - Net primary school enrolment rate - Gross enrolment rate - Drop out rate - Transition rate from primary to secondary - Proportion of students in grade seven passing at specified mark in standard examination	- Actual and budgetary allocation for basic education
C. Vulnerability	- Built capacity to all communities needing safety nets programs	- Established data base for the vulnerable groups - Promoted the production of drought resistant crops in all drought prone areas - Promoted community managed irrigation schemes in all potential irrigation areas
D. Social well-being	- Fully implemented Poverty Reduction Strategy	- Fully implemented local government reform program - Ratio of decided to filed court cases - Average time taken to settle commercial disputes

		- Ratio of actual Court of Appeal sessions to planned sessions - Number of PRS workshops held and composition of committees
3. Achieve and sustain a conducive development environment		
A. Macroeconomic Stability	- Inflation rate	- Fiscal balance - Gross official international reserves - Exchange rate - Current account balance
B. Governance	- Number of budgetary votes managed through IFMs - Expenditure commitments and arrears recorded through IFMs - Spread and magnitude of corruption. - Integrity and transparency in the accounting system. - A governance system that is efficiently and effectively decentralized. - Strengthened professional and cost effectiveness of the public service system. - Improved public service capacity, motivation and performance. - Improved budget management at central and lower levels	- Rolled out the Integrated Financial Management Information System (IFM) to all ministries and sub-treasuries - Developed and approved specific anti-corruption action plans for the ministries of Agriculture and Cooperatives, Education and Culture; Health; and Water; and the CSD based on the National Anti-corruption Strategy. - Developed and approved performance improvement modules for priority sectors. - Timely prepared budgets at all levels. - Institutional pluralism in the delivery of public services

Monitoring gender issues is an integral part of the PRS monitoring system. Health and education indicators such as infant and under five mortality rates, immunization rates, enrolment rates, and transition rates from primary to secondary education will be disaggregated by gender. In addition, the monitoring system includes gender-specific indicators such as the seropositive rate in pregnant women, maternal mortality and the percentage of births attended by trained personnel.

Health and education indicators will also be disaggregated by rural and urban areas and by administrative regions. This is very important given the large geographical variations in social conditions within the country. For example, infant mortality and under-five mortality rates are three times higher in the most deprived region than in the least deprived.

One of the main challenges is to select a manageable number of indicators that provides relevant and sufficient information for assessing the progress of the PRS. In the case of Tanzania, this has been an iterative process. The monitoring system started with approximately 111 aggregate indicators at the national level; currently, it includes around 70. The process of refining the list of indicators will continue as government officials and their counterparts learn which are the most useful indicators and which ones are missing from the list.

Data sources

Calculating reliable baseline figures for the indicators selected was challenging in some cases because recent data were not available. The most recent consumption data to estimate the incidence of poverty come from the 1991/92 Household Budget Survey (HBS). A baseline estimate for 2000 and tentative targets were set by extrapolating the 1991/92 survey results on the basis of population estimates derived from the listing done for the 2000/2001 HBS, but there are methodological problems with these estimates, and they will be revised based on preliminary results of the next HBS.

Other major sources of baseline data include the 1999 Tanzania Reproductive and Child Health Survey (TRCHS) for health and nutrition indicators; administrative data for education indicators;

and the National Accounts and the Economic Survey for macroeconomic indicators (see Table 2). Data from the 2002 Census and the 2000/01 HBS will help validate the reliability of administrative data for school enrolment and update the mortality figures from the 1999 TRCHS. The latter is also important since the TRCHS is an interim Demographic and Health Survey (DHS) using a relatively smaller sample compared to the full survey.

Table 2. Sources of information for onitoring

Indicator type	Baseline source	Follow-up frequency and data source
Poverty Headcount	Preliminary estimates: 1991/92 Household Budget Survey (HBS) Update: 2000/2001 HBS	No additional HBS have been planned during PRSP implementation period
Proxy income indicators for income poverty	2000/2001 HBS 2002 Census	Annual poverty monitoring surveys will measure proxy indicators for income
Macro-economic indicators	National Accounts and the Economic Survey prepared by National Bureau of Statistics and the Planning Commission	Annual updates from same sources
Rural infrastructure	Road Sector reports prepared by the Ministry of Works and the Ministry of Regional and Local Government	Same source, frequency not specified
Health	1999 Tanzania Reproductive and Child Health Survey (Interim DHS)	2002 Census DHS (expected to be held during the implementation period of this PRSP) Health Information System (for annual updates of immunization coverage)
Proportion of districts with an active AIDS awareness campaign	National AIDS Control Programme	Same as baseline, frequency not specified
Nutritional status of children	1999 Tanzania Reproductive and Child Health Survey (new estimate from next DHS)	Community-level monitoring and routine monitoring at health centers Annual poverty monitoring surveys may also include an anthropometric module
Education indicators	Routine data collection system of the Ministry of Education School Mapping	Annual monitoring using administrative data. 2002 Census will provide a cross-check on the administrative data for enrolment
Resource allocation	PER, MTEF and Annual Budget processes	Same source; quarterly review meetings

It is planned that most indicators will be monitored on an annual basis, except for the poverty headcount and some health and nutrition indicators. There are no plans to field a household income or expenditure survey within the next three years. Therefore, as mentioned earlier, the poverty headcount would be substituted by a set of proxy income indicators with baselines calculated on the basis of the 2000/2001 HBS and the 2002 Census, and may be tracked through an annual poverty monitoring survey. The instrument for annual monitoring has not been defined yet; one option is to use the Core Welfare Indicators Questionnaire (CWIQ). Health and nutrition indicators will be monitored at least once within the three year period depending on when the results of the next DHS become available.

Planned institutional arrangements for monitoring and evaluation

The proposed institutional framework for monitoring the PRS is the result of broad consultations among different stakeholders. First, proposals were put forward at a stakeholder meeting which included representatives of government, civil society, NGOs, private sector and academic and research institutions. These proposals were discussed at a subsequent meeting attended by

officials from multilateral and bilateral organizations, the Ministry of Finance, the Planning Commission and the National Bureau of Statistics (NBS). The meeting was organized by the Vice-President's Office in its role as coordinator of the PRSP preparation.

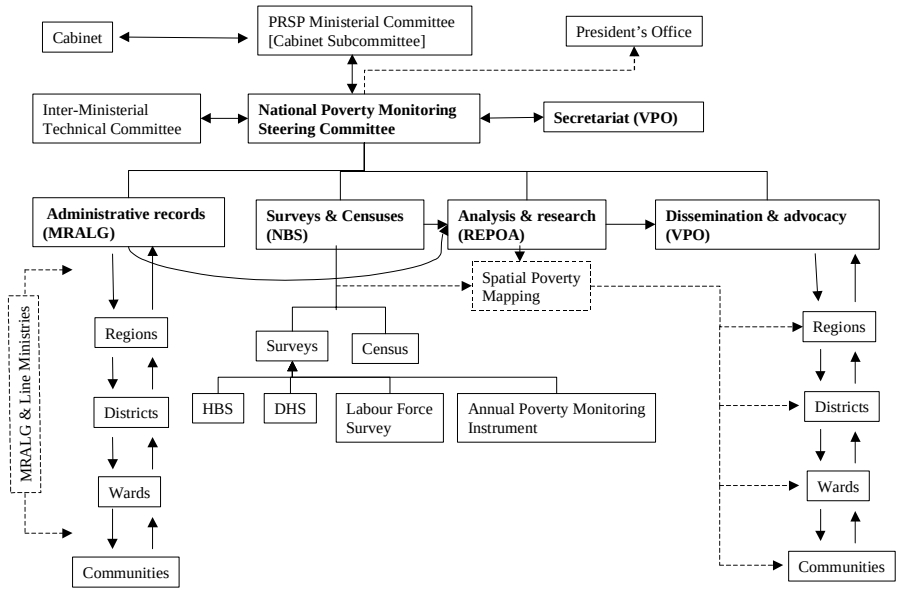
The envisioned apex of the M&E system is the National Poverty Monitoring Steering Committee (NPMSC). Its role would be to provide overall guidance on PRS monitoring and ensure that feedback from the monitoring system gets incorporated into national policy making. The Committee would be integrated by representatives of the government, private sector, NGOs and civil society. The Vice-President's Office would serve as its secretariat.

As illustrated in Figure 1, the NPMSC would be assisted in its task by four working groups:

- The Surveys and Census working group led by the National Bureau of Statistics will be responsible for conducting large household surveys and the census, and for coordinating data storage activities through the Socio-Economic Database initiative.
- The Routine Data working group led by the Ministry of Regional Administration and Local Government (MRALG) will be responsible for coordinating and managing sectoral data collection from line ministries as well as data collected through the administrative systems of decentralized government units.
- The Research and Analysis working group led by the President's Planning Commission and the Research on Poverty Alleviation (REPOA) group will be responsible for coordinating special studies and initiatives such as the spatial poverty mapping.
- The Dissemination and Advocacy working group led by the Vice-President's Office will be in charge of coordinating dissemination activities at all levels and ensuring that the views of local governments are reflected in the monitoring system.

The National Poverty Monitoring Steering Committee would play a key role as a link between policy makers and the monitoring system. The NPMSC would liaise with the PRSP Ministerial Committee through the Vice-President's Office. This committee, which includes several ministers and the Governor of the Bank of Tanzania, was formed to guide the PRSP preparation process and implementation. It is supported by the Inter-Ministerial Technical Committee coordinated by the Ministry of Finance and comprising officials from the Vice President's Office, Prime Minister's Office, Planning Commission, Bank of Tanzania and several line ministries. (As of March 2001, the NPMSC had not been constituted, but the Working Groups had begun working under the guidance of the PRSP Technical Committee.)

Figure 1 : Institutional Framework for PRSP Monitoring in Tanzania



Source: Aide-Memoire: Tanzania PRSP Partnership Mission, December 2000

The proposed institutional framework for PRS monitoring and evaluation would provide a good link between data producers and users, but also pose a number of challenges. First, strong institutional capacity is needed in the MRALG and the NBS to fulfill their coordination role successfully. Second, the role of the Ministry of Finance as coordinator of public expenditure tracking is not clearly captured in the current framework. It would be important to establish coordination mechanisms between the MOF and the MRALG, which is in charge of coordinating all administrative data for monitoring. Third, full implementation of the local government reform now underway is necessary to ensure an adequate flow of administrative data from different government levels. This reform will clarify the division of responsibilities in managing information systems between the MRALG and line ministries such as Education and Health. Further delays in implementing the reform may result in duplication of efforts or missing information. Finally, the proposed institutional framework lays out a fairly clear structure for monitoring and evaluation at the national level, but arrangements at the regional and district level are less clear. To the extent that decentralization efforts devolve decision making power to local level governments it is essential that a structure for monitoring and evaluation at the local level is in place. Overall, it is important that the incentives for collecting, analyzing and reporting information which is accurate and timely are consistent at all levels of the monitoring system.

3. Sources

Republic of Tanzania. 2000. *Poverty Reduction Strategy Paper*

CS 3. Citizen Feedback Surveys as a Tool for Civil Society Participation in Assessing Public Sector Performance: The Case of Bangalore in India

In Bangalore, India, an NGO has conducted citizen feedback surveys focused on services provided by the municipal government, such as water and electricity, garbage collection, and hospitals. Citizens are asked whether they are satisfied with these public services, which aspects are most or least satisfactory, whether government staff are helpful, and whether bribes have to be paid to officials to obtain these services.

The objectives of the survey are:

- To generate citizen feedback on public services and give each municipal agency an overall grade on its performance;
- To identify which specific services are delivered well, or poorly;
- To identify the breadth and depth of corruption;
- To catalyze citizens to be more proactive;
- To provide a diagnostic tool for the municipal departments so that their senior management can better understand their agencies' performance and identify aspects of the services where performance can be improved; and
- To encourage and prod public agencies to be more client-oriented and transparent.

The Bangalore surveys have ranked all municipal government agencies on the basis of the level of citizen satisfaction with their delivery of services. Hospitals and banks received high ratings; the city development authority— with the highest levels of reported corruption—received the lowest rating.

The results of the surveys have been widely published, with lively press coverage. Workshops have been held to provide the findings to citizen groups and other NGOs. Although the findings were not news to them, they provided hard evidence and allowed specific problem areas to be pinpointed. The findings have also stimulated citizen participation and the formation of residents' groups.

The NGO that conducted the surveys gave detailed reports to the heads of all government service agencies. Most agency heads and senior officials were lukewarm to the findings, but some responded well, such as the head of the city development authority, who subsequently initiated a partnership approach with citizen groups and NGOs. This led to innovations in service delivery and a new system for airing client grievances. With the NGO's help, training programs for officials and a partnership group to disseminate information and act as a watchdog were set up.

Similar surveys have now been conducted for other cities in India, including Madras, Mumbai, Calcutta, and Pune. This has enabled comparisons for a number of cities to be published.

Source: Adapted from MacKay, Keith, and Sulley Gariba (eds.). 2000. *The Role of Civil Society in Assessing Public Sector Performance in Ghana: Proceedings of a Workshop*. Evaluation Capacity Development, Operations Evaluation Department, World Bank, Washington, D.C.

CS 4. Evaluating the gains to the poor from workfare: Argentina's Trabajar Program

1. 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 an 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. (For more information on this and other public works programs see the chapter on **Social Protection**). The infrastructure projects 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 as a result of the evaluation results. Trabajar 2 included a number of reforms designed to improve project targeting. The central government's budget allocation criteria gave increased weight to provincial poverty and unemployment indicators, and 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 reforms could increase 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 demonstrate 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.

2. 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. It 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. Commonly available data are exploited by using a new methodology for assessing poverty targeting when there is no actual data on program spending incidence.

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.

3. Data Collection and Analysis Techniques

The assessment of net income gains to program participants draws on two data sources, a national living standards survey (Encuesta de Desarrollo Social – EDS) and a survey of Trabajar participants conducted specifically for the purposes of evaluation[3]. 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 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).

To generate the matching control group from the EDS survey, the study uses a technique called propensity scoring[4]. An ideal match would be two individuals, one in the participant sample and one in the control group, for whom all of the 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)[5]. 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 did not entail any special data collection. It draws on data from the Ministry's project office on funding allocations by geographic department. It also draws on a poverty index for each department, calculated from the 1991 census as the proportion of households with 'Unmet Basic Needs' (UBN)[6]. 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 consisted of a two-stage study of Trabajar infrastructure projects. In the first stage a sample of 50 completed Trabajar 1 projects were rated based on indicators in six categories: technical, institutional, environmental, socioeconomic, supervision, and operations and maintenance. Projects were then given an overall quality rating according to a point system, and cost-benefit analyses were performed where appropriate (not for schools or health centers). A

[3] 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.

[4] The EDS questionnaire is very comprehensive, collecting detailed data on household characteristics which helps predict program participation, and facilitates the use of the propensity scoring technique.

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

[6] 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).

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.

4. Results

Program impact 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[7]. Nonetheless, 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.

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. Policy Implications

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 efforts by the project team to enhance operating procedures – insisting on more site visits for evaluation and supervision, penalizing agencies with poor performance at project completion, and strengthening the evaluation manual.

6. Evaluation Costs and Administration

[7] 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.

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 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.

7. 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.

8. Sources and further reading

Jalan, Jyotsna, and Martin Ravallion. 1999. "Income Gains to the Poor from Workfare: Estimates for Argentina's Trabajar Program." Policy Research Working Paper 2149. World Bank, Development Economics Research Group, Washington, D.C.

Ravallion, Martin. 1999. "Monitoring Targeting Performance when Decentralized Allocations to the Poor are Unobserved." Policy Research Working Paper 2080. World Bank, Development Economics Research Group, Washington, D.C.

CS 5. Evaluating Bolivia's Social Investment Fund

1. Introduction

Project Description: The Bolivian Social Investment Fund (SIF) was established in 1991 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 America, Africa and Asia. For more information on social investment Funds see the chapter on **Social Protection**.

2. 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.

3. Data Collection and Analysis Techniques

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

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.

4. 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 6th 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.

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.

5. Policy Implications

There is 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.

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

7. Evaluation Costs and 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, 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.

8. Source

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

CS 6. Impact of Active Labor Programs: Czech Republic

1. Introduction

Project Description Many developing and transition countries face the problem of retraining workers when state-owned enterprises are downsized. Retraining programs differ in nature and effectiveness – some are simply disguised severance pay for displaced workers; others are disguised unemployment programs. Hence the importance of evaluating such programs.

Training programs are particularly difficult to evaluate, however. Typically several different programs may be 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.

2. Research Questions and Evaluation Design

The evaluation was part of a broader evaluation of four countries: the Czech Republic, Poland, Hungary and Turkey. 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.)

The evaluation focussed on two questions: first, are participants in different ALPs are more successful at re-entering the labor market than are non-participants and does this vary across subgroups and with labor market conditions; and second, what is the cost-effectiveness of each ALP and how can it be improved.

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. Specifically, 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.

3. Data

This evaluation used government administrative data to create the sample frame for the survey. Twenty 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.

The survey was piloted in four districts. This not only identified technical problems, but also a legal problem that can often arise with the use of administrative records: 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.

4. 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.

5. 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 The evaluation did usefully identify one program which appeared to have wasted money – no impact was shown either overall or for any subgroup. The presentation of the evaluation itself – to be read by policy makers was useful providing tables for each program summarizing the combined benefits in terms of wages and employment –both in aggregate and for each subgroup.

A negative point is that 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.

6. Lessons Learned

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. Moreover, a very important practical lesson is 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 meant that the survey instrument was severely compromised.

A second set of lessons relates to how to structure the analysis so as to provide policy relevant information. This was made possible by a detailed evaluation of the program impact by subgroup. This evaluation 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.

7. Source

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

Burtless, Gary. 1995. “The case for randomized field trials in economic and policy research.” *Journal of Economic Perspectives* 9(2):63-84.

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

CS 7. Impact of Credit with Education on Mothers' and Their Young Children's Nutrition: Lower Pra Rural Bank Program in Ghana

1. Introduction

Project Description The Credit with Education program combines elements of a Grameen Bank –type program with education on basic health, nutrition, fertility and small business skills. The aim is to improve the nutritional status and food security of poor households in Ghana. A partnership was formed between international NGOs and 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.

2. Research Questions and Evaluation Design

The research questions focussed on the program's effects on:

- the nutritional status of children
- women's economic capacity (income, savings time) to invest in food and health care
- women's knowledge and adoption of breast-feeding, weaning, and diarrhea management and prevention practices
- 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 communities were divided 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.

3. 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.

4. 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.

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 undermines the potential for examining the policy implications of the findings. 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.

5. 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.

6. 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 and they 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.

7. 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.

8. Source and Further Reading

MkNelly, Barbara and Christopher Dunford (in collaboration with the Program in International Nutrition, University of California, Davis). 1998. "Impact of Credit with Education on Mothers' and their Young Children's Nutrition: Lower Pra Rural Bank Credit with Education Program in Ghana." Research Paper 4. Freedom from Hunger, Davis, CA.

CS 8. Evaluating Kenya's Agricultural Extension Project

1. 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 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.

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.

2. 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.[8] It draws on both quantitative and qualitative methods so that rigorous empirical findings on program impact could be complemented with beneficiary assessments and staff interviews that highlight practical issues in the implementation process. The study also applied the contingent valuation method to elicit farmers' willingness to pay for extension services[9]. 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.

[8] 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.

[9] 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.

3. 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.[10] 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 agents and farmers drawn from the 1998 OED survey. The survey data are also used to assess 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.

4. Results

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.

5. Policy Implications

[10] 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.

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. For example, 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. The evaluation findings also 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

6. Evaluation Costs and Administration

Costs: The total budget allocated for the evaluation was $250,000, which covered household survey data collection and processing ($65,000 – though this is 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 World Bank staff.

7. 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.

8. Source

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

CS 9. Evaluating Nicaragua's School Reform: A Combined Quantitative-Qualitative Approach

1. 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 design is innovative in that it combines both qualitative and quantitative assessment methods. 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 determine 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.

2. 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).

3. Data Collection and 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[11]. 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 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).

4. 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

[11] 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.

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[12]. 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, and so on.

5. 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.

6. Evaluation Costs and Administration

Costs: The total cost of the evaluation was approximately $495,000, representing less than 1.5% of the World Bank credit[13]. 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, the World Bank and researchers from the Harvard School of Education.

7. 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 turn facilitated rapid capacity building and credibility for the evaluation process within the Ministry.

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

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

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.

8. Sources and Further Reading

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. 5. World Bank, Development Economics Research Group, Washington, D.C.

King, Elizabeth, and Berk Ozler. 1998. “What’s Decentralization Got to Do with Learning? The Case of Nicaragua’s School Autonomy Reform.” Working Paper Series on Impact Evaluation of Education Reforms 9. World Bank, Development Economics Research Group, Washington, D.C.

King, Elizabeth, Berk Ozler and Laura Rawlings. 1999. “Nicaragua’s School Autonomy Reform: Fact or Fiction?” Working Paper Series on Impact Evaluation of Education Reforms 19. World Bank, Development Economics Research Group, Washington, D.C.

Nicaragua Reform Evaluation Team. 1996. “Nicaragua’s School Autonomy Reform: A First Look.” Working Paper Series on Impact Evaluation of Education Reforms 1. World Bank, Poverty and Human Resources Division, Policy Research Department, Washington, D.C.

Nicaragua Reform Evaluation Team. 1996. “1995 and 1997 Questionnaires, Nicaragua School Autonomy Reform.” Working Paper Series on Impact Evaluation of Education Reforms 7. World Bank, Development Economics Research Group, Washington, D.C.

Rawlings, Laura B. Forthcoming. "Evaluating Nicaragua's School-Based Management Reform." In Michael Bamberger, ed., *Integrating Quantitative and Qualitative Methods in Development Research.* World Bank, Poverty Reduction and Economic Management Network, Gender Division, Washington, D.C.

CS 10. The Impact of Alternative Cost Recovery Schemes on Access and Equity in Niger

1. 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).

2. Research Questions and Evaluation Design

The main questions were the impact of the treatment on:

- the demand for and utilization of public health care facilities
- specific target groups (poor, women, and children)
- financial and geographic access
- the use of alternative services
- 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.

3. 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.

4. Econometric Techniques

The study combines comparisons of means with simple logit techniques. Specifically, logit techniques are used to address the issue of utilization patterns; the effect on subgroups; and the effect of geographic and financial access. 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.

5. 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 Abt Associates.

6. Results

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.

7. 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.

8. Source

Diop, F, A Yazbeck and R. Bitran. 1995. "The impact of alternative cost recovery schemes on access and equity in Niger." *Health Policy and Planning* 10(3): 223-40

Wouters, A. 1995. "Improving quality through cost recovery in Niger." *Health Policy and Planning* 10(3): 257-70

CS 11. Schooling Outcomes in Philippine Elementary Schools: Evaluation of the Impact of Four Experiments

1. 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 a quarter of Philippine children drop out before completing sixth grade and those who leave have often mastered less than half of what they have been taught. The government embarked on a Dropout Intervention Program (DIP) in 1990-92 to address these issues. Four experiments were undertaken: provision of 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.

2. Research Questions and Evaluation Design

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

- Two districts in each of five regions of the country were identified as a low-income municipality. In one district the treatment choices were packaged as control, MLM or MLM-PTP; in the other control, SL of SL-PTP. The assignment of the two intervention packages was by a coin flip
- 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
- 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

3. Data

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.

4. 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 (AP_{ist}), which the authors model as:

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

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.

5. 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.

6. 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.

7. 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.

8. Source

Tan, J.P., J. Lane and G. Lassibille. 1999. **"**Schooling Outcomes in Philippine Elementary Schools: Evaluation of the Impact of Four Experiments." *World Bank Economic Review* 13(3):

CS 12. Assessing the Poverty Impact of Rural Roads Projects in Vietnam

1. Introduction

Project Description: Rural roads are being extensively championed by the World Bank and other donors as poverty alleviation instruments. The Vietnam 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 Vietnam 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 Vietnam as a result of the ongoing transition to a market economy. The evaluation began concurrent with project preparation, in early 1997, and is in process. No results are available yet. The evaluation is compelling in that it is one of the first comprehensive attempts to assess the impact of a rural roads project on welfare outcomes – the bottom line in terms of assessing whether projects really do reduce poverty. The design attempts to improve 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.

2. Evaluation Design

The design of the Vietnam 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.

3. Data Collection and Analysis Techniques

Data collected for the purposes of the evaluation include commune- and household-level surveys, along with district-, province- and project-level databases. The **baseline and follow-up commune and household surveys** were conducted in 1997 and 1999, and third and fourth survey rounds, conducted at two-year intervals, are planned. The survey sample includes 100 project and 100 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[14]. 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[15].

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 Vietnam 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, 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 Vietnam'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.

[14] 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.

[15] 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).

4. Evaluation Costs and 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. 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.

5. Source

van de Walle, Dominique. 1999. "Assessing the Poverty Impact of Rural Road Projects." World Bank, Development Economics Research Group, Washington, D.C.