Annex C
Monitoring and Evaluation: Technical Notes and Case Studies

Tables

Figure

Box

Technical Note C.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, intended to improve public actions. Different types of evaluation address different evaluation questions. These questions can be broadly classified in three categories:

- ***Process*** questions aim to understand how the program or a specific component of it are being implemented as originally designed.
- ***Outcome*** questions seek to assess whether the situation of individuals or households in terms of key outcomes (knowledge, behavior, well-being, and so on) has changed, and the extent to which the program is responsible for the observed changes. Outcomes may change for a number of reasons, many of which may be independent of the program. Attribution questions ask whether observed changes were caused by the program or whether they would have occurred anyway.
- Questions about ***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 (for example, households, individuals, facilities, communities, and so on) and how they are going to be selected (opportunistically or using systematic sampling strategies); which kind of comparisons will be made (for example, no comparison, comparison across time or space, comparison of different groups, and so on); and the timing of the data collection (for example, before and after the program, immediately after the program only, during program implementation, and so forth).

Process evaluations assess how effectively a public action is being implemented; they focus on aspects such as who is participating, what activities are being offered, what actions have been taken, and what staff practices and client responses are. A process evaluation may be conducted when problems such as delays, cost overruns, or beneficiary dissatisfaction have been detected by the monitoring system, or may be carried out regularly 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 judgment.

Outcome evaluations assess what happened to individuals (or other units of analysis) after policy or program implementation; they focus on intervention outcomes such as whether people are healthier, better educated, or less vulnerable to adverse shocks. Evaluation designs for outcome evaluations vary along a continuum of levels of complexity. At 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 at this end of the spectrum tend to be relatively simple and quick to yield results, but they leave room for differing interpretations of how much of a change can be attributed to a particular intervention. This type of evaluation generally looks only at the group of program participants; there is no comparison with people or communities that did not participate in the program nor any effort to isolate program or policy effects from other events occurring simultaneously. The evaluation can look at outcomes either 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 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 C.2 and C.3). These evaluations are usually referred to 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 or 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 links between inputs, activities, and outcomes and aim to determine whether a breakdown has occurred—and if so, where, why, and how. They present an 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 failed.

Technical Note C.2 Impact Evaluation Designs

Experimental or randomized designs involve gathering a set of individuals (or other units 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, scheduled to receive funding at a later stage, can be used as control group. Since program participants are selected randomly, any difference from nonprogram participants is due to chance. For this reason, experimental designs are usually regarded as the most reliable method and the one yielding results that are easiest to interpret. 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 study C.7).

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 nonparticipants 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 C.4). Evaluations using matching methods are often easier and cheaper to implement than experimental designs, but the results are less reliable and interpreting them 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 C.5). There is, however, a major drawback with this method: the situation of program participants before and after the intervention may change owing to 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 the result of 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.

Technical Note C.3 Impact Evaluation Methods for Policies and Full-Coverage Programs

When policies or programs affect the whole population, it is generally not possible to identify or construct a control group, and assessing whether such interventions caused changes in outcomes is considerably more difficult. Several methods can be employed.

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 built on data from national accounts, household expenditure surveys, and other survey data. CGE models simulate the counterfactual, though the strength of the model is entirely dependent on the quality of the underlying data and the validity of the assumptions. This can be problematic, as databases are often incomplete and many of the parameters needed cannot be estimated by formal econometric methods. CGE models are also very time consuming, cumbersome, and expensive to develop.

"With and without" comparisons compare the behavior of key variables in a sample of program countries or regions to their behavior in nonprogram countries (a comparison group). Thus this method uses the experiences of the nonprogram countries 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 nonprogram areas and that external factors either affect both groups equally or their impact can be identified separately from that of the intervention.

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

Source: Baker, Judy. 2000. *Evaluating the Poverty Impact of Projects: A Handbook for Practitioners*. Directions in Development. Washington, D.C.: World Bank.

Technical Note C.4 Types of Data Sources for Impact Evaluation

Longitudinal or panel datasets include information on the same individuals (or other units of analysis) for at least two different points in time, one before the intervention (the baseline) and another afterward. Panel datasets are highly valued for program evaluation, but they can be expensive and require substantial institutional capacity (see case study C.5 and chapter 1, "Poverty Measurement and Analysis").

Cross-section data collect information from different people at different points in time. Evaluations using cross-section data usually cost less than studies using information from more than one point in time but, since it is often difficult to tell whether changes were due to the intervention or to other factors, the results tend to be less reliable, except for experimental designs (see case study C.4).

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 trends pre- and postprogram. However, many data points before and after the program are required for rigorous analysis. Time series are used primarily to evaluate policies and programs with full or national coverage.

Case Studies

Case studies C.1 and C.2 provide examples of national poverty monitoring systems, whereas case study C.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 C.4 to C.7 (adapted from Judy Baker [2000], *Evaluating the Poverty Impact of Projects: A Handbook for Practitioners*, World Bank, Washington, D.C.) 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 C.1). Each study includes a discussion on the relative strengths and weaknesses of each evaluation.

Case Study C.1 Monitoring the Progress of the Poverty Eradication Action Plan in Uganda[1]

C.1.1. Introduction

In 1995, the government of Uganda embarked on the formulation of the Poverty Eradication Action Plan (PEAP) to ensure that poverty reduction was the focus of its overall growth and development strategy. This plan was developed 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 proportion of the population living in absolute poverty from 56 percent (1992) to 10 percent and cutting the proportion of people living in relative poverty from more than 85 percent to 30 percent by 2017.

Additional goals were agreed on in four areas—macroeconomics, governance, income generation, and human development—and expanded into a set of strategic objectives (see box C.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.

C.1.2. 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 toward a decentralized, participatory monitoring system with a clearer delineation of roles and responsibilities, including mechanisms for collaboration across institutions.

Box C.1. Poverty Eradication Action Plan Goals, Uganda

Overarching goal: To reduce the proportion of the population living in absolute poverty from 56 percent to 10 percent and to cut the proportion of people living in relative poverty from more than percent to 30 percent by 2017.

Goal #1. Implementation of macroeconomic 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**
- **Refocus 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 and 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**
- **Enhance labor productivity, giving priority to employment of women, reduction of the exploitation of child labor, and safeguarding of 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 the cost-effectiveness of water provision**
- **Achieve universal primary education and improve the quality of education**
- **Promote access to basic education for vulnerable children (for instance, the homeless and street children)**
- **Promote the acquisition, use, and retention of functional literacy by all the people of Uganda**

The system consists of three core institutions:

- 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 policymakers. 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.

Table C.1. Summary of Impact Evaluation Case Studies

Program/ project	*Country*	*Database type*	*Unit of analysis*	*Outcome measures*	*Econometric approach*				*Qualitative evaluation*	*Strengths*
					Random-ization	*Matching*	*Reflexive compari-sons*	*Instru-mental variables*		
Education										
School autonomy reform	Nicaragua	Panel survey and qualitative assess-ments	Students, parents, teachers, directors	Test scores, degree of local deci-sionmaking	No	Yes	Yes	No	Yes	Qualitative-quantitative mix
Dropout intervention	Philippines	Baseline and post intervention survey	Students, class-rooms, teachers	Test scores and drop-out status	Yes	No	Yes	Yes	No	Cost/bene-fit analysis; capacity building
Labor Programs										
Trabajar program	Argentina	Household survey, census, administra-tive rec-ords, social assess-ments	Workers, households	Income, targeting, costs	No	Yes	No	Yes	Yes	Judicious use of ex-isting data sources, innovative analytic techniques
Agriculture										
National extension project	Kenya	Panel data, beneficiary assess-ments	House-holds, farms	Farm pro-ductivity and effi-ciency	No	No	Yes	No	No	Policy-rele-vance of results

Source: Adapted from Baker, Judy. 2000. *Evaluating the Poverty Impact of Projects: A Handbook for Practitioners.* Directions in Development. Washington, D.C.: World Bank.

In addition, other institutions such as nongovernmental organizations (NGOs), academic institutions, research centers, and donors play important, but not yet systematic, roles in collecting and analyzing additional data. Policymakers 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 decisionmaking).

The system is undergoing a major revision aimed at:

- Increasing participation, that is, promoting greater involvement in monitoring activities at the local level and collaboration among the UBoS, the Poverty Monitoring Unit, NGOs, 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 for data analysis and dissemination at central levels, in order to decrease the lag time between data collection and analysis/dissemination.
- Defining institutional roles, that is, setting clearly defined roles and responsibilities, including mechanisms of collaboration.
- Harmonizing progress reporting, that is, 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 fieldwork; actual data collection and data entry; and carrying out all other functions associated with fieldwork. 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.

The Ministry of Finance, Planning, and Economic Development prepared a Poverty Monitoring and Evaluation Strategy in October 2001 to discuss some of these issues and possible solutions.

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 not focused on priorities. 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 table C.2). This list will be further refined 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. Educational data are the only data that are disaggregated by sex. This is a major limitation for a complete poverty analysis and is expected to be addressed in the future.

Although some progress has been made in aligning the indicators with the goals of the PEAP, 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 is that some indicators are not unambiguous measures of progress—that is, 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 may be spending more on education because they consume more or because they have to pay more and may, in fact, be consuming less. Finally, the list does not distinguish between final and intermediate indicators, a distinction that would be useful when judging overall progress.

Table C.2. Revised List of Monitoring Indicators, Uganda

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 share of food expenditure	National, regional
Proportion of population living under thatched roofs	National, regional
Dependency ratio	National, regional, district
Gini coefficient	National, rural/urban
Consumption per capita of poorest 20 percent	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 resulting from insurgency	National, regional
Number of criminal cases reported	National, regional
Proportion of households experiencing major income shocks last year	National, regional
Refugees and displaced persons as proportion of district population	District
Proportion of households under economic distress selling assets	National
Road network	
Road length opened	National
Road length upgraded	National
Proportion of districts with more than 50 percent 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 microfinance portfolio	National, district
Proportion of population accessing microcredit	National, district
Growth in savings	National, district
Credit management (effective use)	National, district
Availability of microfinance services	National, urban/rural

Table C.2. Revised List of Monitoring Indicators, Uganda *(continued)*

Indicators	*Intended level of disaggregation*
QUALITY OF LIFE	
Health Indicators	
Incidence of disease	National, district
Immunization coverage	National, district
Proportion of population within 5 km. of 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 .5 km of 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 four 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	
Level of compliance with environmental standards	All National
Corrective actions by the National Environmental Management Agency	
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 about rights/entitlements	National
Proportion of reported cases cleared	National
Number of people on remand beyond the specified period by law	National
Number of backlogged 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 in 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 districts can use for planning, implementing, and monitoring their programs and policies. UBoS and the Ministry of Local Government are working on a system to involve the District Planning Units in data collection to ensure that relevant statistics and qualitative information are used 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; and the 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 2001/2002 and a new round is planned for 2003/2004.

Monitoring Surveys, which collect information similar to that collected by the IHS but use a smaller sample of 5,000 households and a shorter questionnaire (which does include a consumption module). They have been conducted annually from 1992–93 to 1997.

Demographic and Health Surveys (DHS) collect information on maternal and child health, immunization, health care access, major disease incidence, and so forth. Baseline data were provided by the 1995 DHS and a follow-up survey was conducted in 2001.

Also, the Population Census of 2002 will provide updated information on the demographic structure of the population: age, marital status, ethnicity, religion, household size, dependency ratios, and so forth. Other surveys, such as the Public Expenditure Tracking Survey (see chapter 6, "Public Spending") 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. This survey was piloted in 1996 and conducted nationwide in 1999 (currently by the Ministry of Public Services but in the future by the Bureau of Statistics). A survey of health facilities was recently completed.[2]

Management information systems (MIS) collect sectoral information on outputs, access to services, and, to a limited extent, 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 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 collecting district-level information on enrollment of pupils, number of teachers, teaching/learning materials, facilities, and finances.

A number of problems with the MIS data have been identified. First, information is incomplete. By 1996, for example, the education census had a response rate of 60 percent 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 larger grants. Random checks that have been implemented reveal enrollment overreporting. In the health sector, diagnostic tools, staff capacity, and communication infrastructure are limited in many areas—especially remote rural areas—so that gross underreporting of disease incidence occurs. Third, there is an issue of timeliness; the arrival of data from districts is slow, and data analysis, compilation, and reporting at the center are delayed. The Education Statistics Abstracts, for example, are usually produced 1.5 years after data collection. So the data are not used for service provision and planning.

One reason identified for the poor performance of MISs 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 ministries. District-level activities include 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.

As for ***qualitative studies***, 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. Field work for the latest participatory poverty assessment was conducted in late 2001. The UPPAP is a partnership of the government, donors, the nine district authorities in which the project operates, and Oxfam, the implementing agency.[3] Within the government, the UPPAP is situated in 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. Despite the large proportion of poor people engaged in agricultural activities, agriculture data are not readily available. The Program on Modernization of Agriculture, a central element of the third pillar of the PEAP, does not define indicators, and the District Resource Information System does not include agriculture data. Likewise, there is little usable information on governance issues except in the National Integrity Survey

and reports from the Human Rights Commission. A wealth of data has been collected by the Inspectorate General of Governance, but because this organization is understaffed, no summary statistics are available. Indicators for the PEAP in this area have not been defined.

Data analysis, dissemination, and feedback

Data analysis is conducted mostly at the central level by UBoS. At the district level, data analysis is limited. A few districts have started their own monitoring systems under the Local Government Development Project, the main objective of which 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 to 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 toward 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. The next report is planned for 2003.

Reports are disseminated at the national and district levels and are used in the revision of the PEAP, the MTEF, and sector reviews. The Poverty Monitoring Unit and the Poverty Working Group (PWG) ensure that the monitoring results are used in policymaking and budget allocation processes. Specifically, the PWG, which includes government officials and 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. The PWG also makes recommendations on the overall budget allocation of resources for poverty reduction as well as on other budget policies that affect the poor. Despite these efforts, the performance of public expenditures is still mostly measured in terms of inputs and activities rather than contributions to poverty reduction. Progress toward goals still plays a limited role in the sectoral budget allocation process. An incentive system linking resource allocation and performance assessment to contributions to PEAP outcomes needs to be developed.

C.1.3 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 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 facilitates 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 macroeconomic data and output tables, which are then used as input sources of reports, newsletters, or for electronic dissemination.

Another area of emphasis 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 mobile teams. These teams will be used both to conduct UBoS surveys and 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 survey that could be carried out by district governments to meet their information needs.

The program also supports the repeated administration of an annual National Service Delivery Survey. 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. The survey will be progressively mainstreamed and taken over by UBoS. Future rounds of the survey will be supplemented with focus groups interviews.

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 Finance, Planning, and Economic Development, Kampala.

———. 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

World Bank. 1999. "Uganda's Integrated Information Management System: A New Approach in Statistical Capacity-Building." Africa Region Findings, No. 142, Washington, D.C.[4]

Case Study C.2 Proposed Plan to Monitor the Poverty Reduction Strategy in Tanzania

C.2.1 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, 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
- achieving and sustaining an environment conducive to development.

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.

C.2.2 Monitoring the Poverty Reduction Strategy

The Poverty Reduction Strategy Paper (PRSP) 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 Tanzanian 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 monitoring and evaluation (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; these will be used to monitor progress toward the main goals of the strategy. Intermediate indicators will be used to monitor 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 objective of the PRS is to reduce income poverty. Thus it is important to monitor at regular intervals the proportion of the population living below the poverty line. However, in the case of Tanzania, as in many other countries, collecting income or expenditure data at frequent intervals is 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 C.3, 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 Millennium Development Goals: halving the incidence of poverty 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; a survey module is being developed to collect this information on a yearly basis. The intermediate indicators chosen are relevant in the Tanzanian context and should also be available annually 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 first- and second-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, breastfeeding practices), since the main data source for these indicators will be the Demographic and Health Survey that is expected to be conducted during the implementation period of the PRS.

Education indicators and targets are relevant to the goal of eradicating illiteracy by 2010: raise gross primary enrollment to 85 percent, increase the transition rate from primary to secondary school from 15 to 21 percent, reduce the dropout rate in primary school from 6.6 to 3 percent, and raise net primary school enrolment from 57 to 70 percent. Most of the final indicators are clearly defined, except for gender equity, which could be measured with respect to gross enrollment rates, net enrollment 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, and average travel time to school, could supply information useful for understanding 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. Enrollment 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 they monitor activities rather than results; no "final indicator" is included. More specific intermediate indicators would also 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 a database 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 they will give only a very partial picture of progress on these issues. 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 gross domestic product (GDP) annual growth

Table C.3. Proposed Indicators for Monitoring the PRSP, Tanzania

Objectives	*Final indicators*	*Intermediate indicators*
1. Reducing income poverty	• Poverty incidence • Ownership of household assets (proxy indicator) • Type of construction materials of dwelling units; e.g., floors, walls, and roofing (proxy indicator)	• 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 two 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 enrollment rate • Gross enrollment rate • Dropout rate • Transition rate from primary to secondary school • Proportion of students in grade seven passing at specified mark in standard examination	• Actual and budgetary allocation for basic education
C. Vulnerability	• Capacity built in all communities needing safety nets programs	• Database for the vulnerable groups established • Production of drought-resistant crops in all drought-prone areas promoted • Community-managed irrigation schemes promoted in all potential irrigation areas
D. Social well-being	• Poverty Reduction Strategy fully implemented	• Local government reform program fully implemented • 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, attendance, and composition of committees • Dissemination of reports

Table C.3. Proposed Indicators for Monitoring the PRSP, Tanzania *(continued)*

Objectives	*Final indicators*	*Intermediate indicators*
3. Achieve and sustain an environment conducive to development		
A. Macroeconomic stability	• Inflation rate	• Fiscal balance • Gross official international reserves • Exchange rate • Current account balance
B. Governance	• Number of budgetary votes managed through Integrated Financial Management Information Systems (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 effectiveness and cost-effectiveness of the public service system • Improved public service capacity, motivation, and performance • Improved budget management at central and lower levels	• IFM rolled out to all ministries and sub-treasuries • Specific anticorruption action plans developed and approved for the Ministries of Agriculture and Cooperatives, Education and Culture; Health; and Water; and the CSD based on the National Anticorruption Strategy • Performance improvement modules developed and approved for priority sectors • Timely preparation of budgets at all levels. • Institutional pluralism in the delivery of public services

over the next three years; this would set the basis for achieving the medium- and long-term poverty reduction goals. 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.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, enrollment 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

Table C.4. Sources of Information for Monitoring, Tanzania

Indicator type	*Baseline source*	*Follow-up frequency and data source*
Poverty headcount	Preliminary estimates: 1991–92 Household Budget Survey (HBS) Update: 2000–01 HBS	No additional HBS have been planned during PRSP implementation period
Proxy income indicators for income poverty	2000–01 HBS 2002 census	Annual poverty monitoring surveys will measure proxy indicators for income
Macroeconomic 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. The 2002 census will provide a cross-check on the administrative data for enrollment
Resource allocation	PER, MTEF, and Annual Budget processes	Same source; quarterly review meetings

from the listing done for the 2000–01 HBS, but there are methodological problems with these estimates, and they will be revised based on the results of the 2000–01 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 C.4). Data from the 2002 census and the 2000–01 HBS will help validate the reliability of administrative data for school enrollment and update the mortality figures from the 1999 TRCHS. This is an important check, since the TRCHS is an interim DHS using a relatively smaller sample than the full survey does.Most indicators will be monitored annually, except for the poverty headcount and some health and nutrition indicators. As mentioned earlier, the poverty headcount would be substituted by a set of proxy income indicators with baselines calculated on the basis of the 2000–01 HBS and the 2002 census, and may be tracked through an annual poverty monitoring survey. The decision was made not to conduct a specific poverty monitoring survey every year, but rather to develop a special module containing the proxy income indicators, to be included in whatever household survey is to be undertaken in a given year. 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 that included representatives of government, civil society, NGOs, the 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. The meeting was organized by the Vice-President's Office in its role as coordinator of the PRSP preparation. A Poverty Monitoring Master Plan was agreed to in November 2001.

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

As illustrated in figure C.1, the NPMSC is assisted in its task by four working groups, coordinated by the Poverty Eradication Division.

The Surveys and Censuses Working Group, led by the National Bureau of Statistics, is responsible for conducting large household surveys and the census, and for coordinating data storage activities through the Socio-Economic Database initiative.

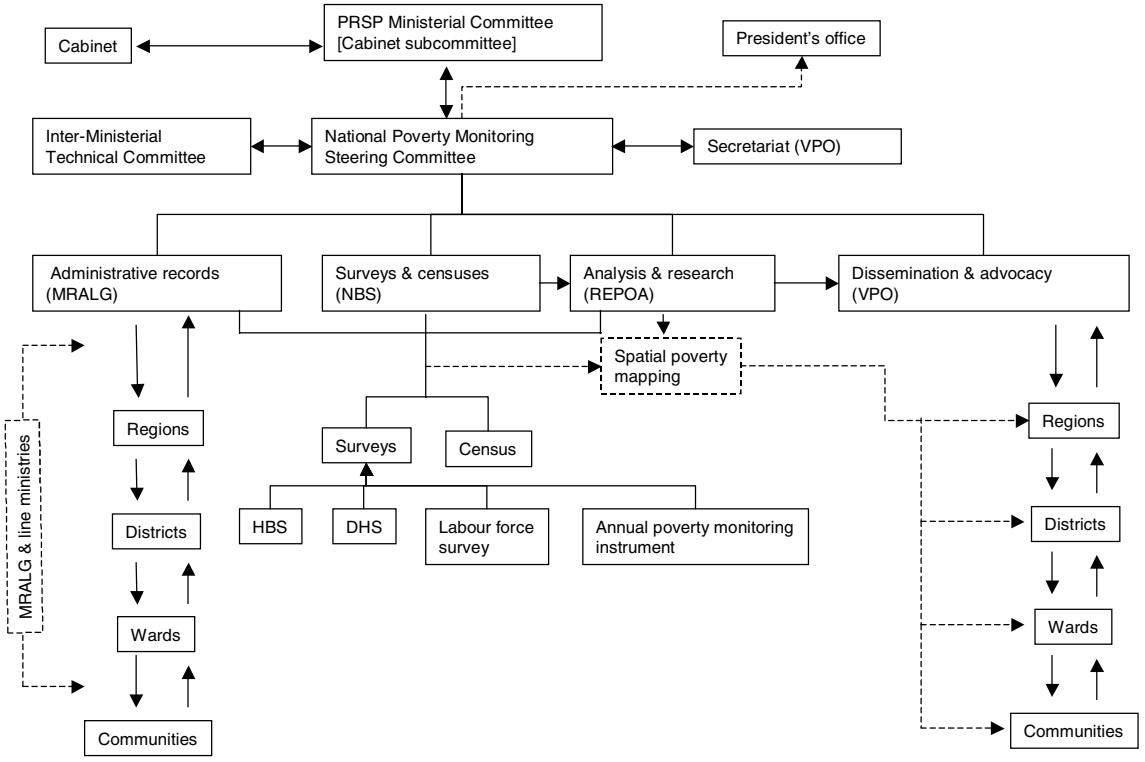
The Routine Data Working Group, led by the Ministry of Regional Administration and Local Government (MRALG), is responsible for coordinating and managing sectoral data collection from line ministries as well as data collected through the administrative systems of decentralized government units.

The Research and Analysis Working Group, led by the President's Planning Commission and the Research on Poverty Alleviation (REPOA) Group, is responsible for coordinating special studies and initiatives such as spatial poverty mapping.

The Dissemination, Sensitization and Advocacy Working Group, led by the Vice-President's Office, is 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 is expected to play a key role as a link between policymakers and the monitoring system, liaising 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, which is coordinated by the Ministry of Finance and comprises officials from the Vice President's Office, the Prime Minister's Office, the Planning Commission, the Bank of Tanzania, and several line ministries.

Figure C.1. Institutional Framework for PRSP Monitoring, Tanzania



The proposed institutional framework for PRS monitoring and evaluation could provide a good link between data producers and users, but it also poses a number of challenges. First, the MRALG and the National Bureau of Statistics must have strong institutional capacities to fulfill their coordination roles successfully. Second, the role of the Ministry of Finance (MOF) 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 under way is necessary to ensure an adequate flow of administrative data from different government levels. This reform will hopefully clarify the division of responsibilities in managing information systems between the MRALG and line ministries such as Education and Health, which at present remains unclear. 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 levels are less clear. To the extent that decentralization efforts devolve decisionmaking power to local level governments, it is essential that a structure for monitoring and evaluation at the local level be in place. Overall, it is important that the incentives for collecting, analyzing, and reporting information that is accurate and timely be consistent at all levels of the monitoring system.

Sources:

United Republic of Tanzania, The. 2000. "Poverty Reduction Strategy Paper." Dar Es Salaam

_____. 2001. "Poverty Monitoring Master Plan." Dar Es Salaam.

Case Study C.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, 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 managers 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 citizens' 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 about the findings, but some responded well, such as the head of the city development authority, who subsequently initiated a partnership approach with citizens' groups and NGOs. This led to innovations in service delivery and a new system for airing client grievances. With the NGOs' 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.

Case Study C.4 Evaluating the Gains to the Poor from Workfare: Argentina's Trabajar Program

C.4.1 Introduction

Argentina's Trabajar program aims to reduce poverty by simultaneously generating employment opportunities for the poor and improving social infrastructure in poor communities. The program offers relatively low wages in order to attract only poor, unemployed workers as participants, who would "self-select" to participate. (For more information on this and other public works programs see chapter 17, "Social Protection.") The infrastructure projects that participants work on are proposed by local governments and NGOs, which must cover the nonwage costs of the project. Projects are approved at the regional level according to central government guidelines.

The program has evolved over time, incorporating lessons learned. Trabajar I, a pilot program, was introduced in 1996 in response to an economic crisis and unemployment rates of more than 17 percent. Trabajar II was launched in 1997 as an expanded and reformed version of the pilot program, and Trabajar III began in 1998. Trabajar II 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; at the local level, efforts were made to strengthen the capacity of provincial offices to help poor areas mount projects and to raise the quality of the infrastructure built under the program.

C.4.2. Evaluation design

The evaluation effort began during the preparation of Trabajar II. The aim of the evaluation was to determine whether or not the program was achieving its goals and to indicate areas where reforms could increase its effectiveness. The evaluation consisted of a number of separate components that assessed: a) the net income gains that accrued to program participants; b) the allocation of program resources across regions (targeting); c) the quality of the infrastructure projects financed. In addition, the study looked at the role of the community and NGOs in project outcome.

Two of the evaluation components demonstrate best practice techniques. The first component of the Trabajar evaluation, the assessment of net income gains, improved upon conventional assessments of workfare programs, which typically measure participants' income gains as simply their *gross* wages earned, by estimating *net* income gains. Drawing upon new and existing household survey data and using recent advances in matched comparison techniques, the study accounted for forgone income (income given up by participants in joining the Trabajar program), which resulted in a more accurate, lower estimate of the net income gains to participants.

The second component, the study of targeting outcomes, introduced a new technique for evaluating targeting (the allocation of program funding) when the incidence of public spending at the local level is unobserved.

The overall evaluation design also presented a best practice mix of components and research techniques—from quantitative analysis to engineering site visits to social assessment—which provided an integrated stream of results in a timely manner.

C.4.3. Data collection and analysis techniques

The assessment of net income gains to program participants drew on two data sources, a national living standards survey (Encuesta de Desarrollo Social, or EDS) and a survey of Trabajar participants conducted specifically for the purposes of evaluation.[5] These surveys were conducted in August and September of 1997 by the national statistical office, using the same questionnaire and interview teams. The sample for the EDS survey covered 85 percent of the national population, omitting some rural areas and very small

communities. The sample for the Trabajar participant survey was drawn from a random sample of Trabajar II projects located within the EDS sample; it generated data for 2,802 current program participants (total Trabajar II participants between May 1997 and January 1998 numbered 65,321).

To generate the matching control group from the EDS survey, the study used a technique called propensity scoring. 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).[6] Data on incomes in the matching control group of nonparticipants allowed the estimation of the income forgone by Trabajar II participants. Net income arising from program participation was then calculated as total program wages minus forgone income.

The targeting analysis remarkably did not entail any special data collection. It drew on data from the ministry's project office on funding allocations by geographic department. It also drew on a poverty index for each department, calculated from the 1991 census as the proportion of households with "Unmet Basic Needs".[7] To analyze targeting incidence, data on public spending by geographic area—in this case, department—were regressed on corresponding geographic poverty rates. The resulting coefficient consistently estimated a "targeting differential" given by the difference between the program's average allocations to the poor and nonpoor. This national targeting differential could then be decomposed into components due to the central government's targeting mechanism (funding allocations across departments) and to targeting done at the provincial level.

The analysis of the quality of infrastructure consisted of a two-stage cost-benefit analysis of Trabajar infrastructure projects. In the first stage a sample of 50 completed Trabajar I projects were given an overall quality rating based on indicators in six categories: technical, institutional, environmental, socioeconomic, supervision, and operations and maintenance, and cost-benefit analyses were performed where appropriate (not for schools or health centers). In the second stage, a follow-up study of 120 Trabajar II projects was conducted a year later, tracking the impact of reforms on infrastructure quality.

Social assessments were conducted during project preparation for both Trabajar I and Trabajar II. They provided feedback on project implementation issues such as the role of NGOs, the availability of technical assistance in project preparation and construction, and the selection of beneficiaries. They were carried out by sociologists and included focus groups and interviews.

C.4.4. Results

Program Impact. Taking account of forgone income is important to getting an accurate picture of workfare program benefits. Program participants could not afford to be unemployed in the absence of the program; hence some income is forgone through program participation. Forgone income is estimated by observing the incomes of nonparticipants "matched" to those of program participants. Matching-method estimates show that ignoring foregone incomes greatly overstates the average gains from the program, which were estimated to be about half the wages received through the program. The evaluation also revealed that the distribution of gains was decidedly pro-poor, with 80 percent of program participants being among the poorest 20 percent of the population, even after reducing income gains by the amount of the forgone income. Female participation in the program is low (15 percent), but net income gains are virtually identical for male and female Trabajar participants.

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

Infrastructure Project Quality. Quality was found to be adequate, but Trabajar II reforms, disappointingly, did not result in significant improvements. Part of the reason was the sharp expansion of the program, which made it difficult to meet some of the operational standards that had been specified ex ante. However, Trabajar II infrastructure 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.

C.4.5. Policy implications

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

C.4.6 Evaluation costs and administration

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

C.4.7. Lessons learned and future work

Importance of accounting for forgone income in assessing the gains to workfare. Forgone income represents about half of the gross wages earned by workfare program participants in Argentina. The results suggest that conventional assessment methods (using only gross wages) substantially overestimate income gains, and hence also overestimate how poor participants would be in absence of the program.

Usefulness of propensity score matching methods. Propensity scores allow reliable matches to be drawn between a participant and a nonparticipant (control group) sample.

Judicious use of existing national data sources. Often, existing data sources such as the national census or household surveys can provide valuable input to evaluation efforts. Drawing on existing sources reduces the need for costly special-purpose data collection. 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 were all used to generate a wide range of results that provided valuable input into the project's effectiveness and pinpointed areas for reform.

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

Future work. Three studies are planned: the matched comparison research technique will be applied again to assess the impact of the program participation on labor market activity; infrastructure project quality will be reassessed; and a qualitative research component will investigate program operations and procedures.

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.

Case Study C.5 Evaluating Kenya's Agricultural Extension Project

C.5.1 Introduction

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 developing institutions 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.

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 particularly strong because of very high estimated returns to T&V reported in an earlier study and the lack of convincingly visible results, including the poor performance of Kenyan agriculture in recent years.

The World Bank's Operations Evaluation Department (OED) undertook an evaluation of the Kenyan extension system in 1997-99. Using a results-based management framework, the evaluation examines the impact of project services on farm productivity and efficiency. It also develops measures of program outputs (for example, frequency and quality of contact) and outcomes (that is, farmer awareness and adoption of new techniques).

C.5.2 Evaluation design

The evaluation strategy illustrates best practice techniques in using a broad array of evaluation methods and exploiting existing data.[8] It drew 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 was complicated by the fact that the T&V system was introduced on a national scale, preventing a "with program" and "without program" (control group) comparison. The evaluation methodology therefore sought to exploit the available preproject household agricultural production data for limited before-and-after comparisons using panel data methods. For this, existing household data were complemented by a fresh survey to form a panel. Beneficiary assessments designed for this study could not be conducted, but the evaluation drew on the relevant findings of two recent beneficiary assessments in Kenya.

C.5.3 Data collection and analysis techniques

The evaluation approach drew on several existing qualitative and quantitative data sources. The quantitative evaluation is based largely on a 1998 household survey conducted by OED. This survey generated panel data by revisiting as many households as could be located from a 1990 household survey conducted by the World Bank's Africa Technical Department, which in turn drew from a subsample of the 1982 Rural Household Budget Survey.[10]

The study evaluated institutional development by drawing on a survey of 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 drew 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, carried out jointly by the government of Kenya, the African Medical and Research Foundation, British Overseas Development Administration, UNICEF, and the World Bank, which inquired about public services, including extension. Quality and quantity of services delivered were assessed using a combination of the findings of participatory (beneficiary) assessments and 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. Contingent valuation methods were used to directly elicit the farmers' willingness to pay for extension services.

The survey data were also used to assess program outcomes, measured in terms of farmer awareness and adoption of extension recommendations. The program's results—its actual effects on agricultural production in Kenya—were 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 were estimated using the Data Envelopment Analysis, a nonparametric technique, to measure changes in farmer efficiency and productivity over time, along with econometric analysis measuring the impact of the supply of extension services on farm production.

C.5.4 Results

The institutional development of NEP I and NEP II has been limited. After 15 years, the effectiveness of extension services has improved little. Although there has been healthy rethinking of extension approaches recently, overall the extension program has lacked the strategic vision for future development. Management of the system continues to be weak, and information systems are virtually nonexistent. The quality and quantity of service provision are poor. Beneficiaries and extension service staff alike report that visits are infrequent and ineffective. Although there continues to be unmet demand for technically useful services, the focus of the public extension service has remained on simple and basic agronomic messages. Yet the approach taken—a high intensity of contact with a limited number of farmers—is suited to deliver more technical information. The result has been a costly and inefficient service delivery system. The analysis showed that extension activities had little influence on awareness and adoption of recommendations, indicating limited potential for impact. In terms of actual impact on agricultural production and efficiency, the data indicated a small positive impact of extension services on the ability of farmers to get the most production from available resources (technical efficiency). However, no effect was found on allocative efficiency (the use of resources given market prices) or overall economic efficiency (the combination of technical and allocative efficiency). Further, no significant impact of the supply of extension services on productivity at the farm level could be established. The data did show, however, that the impact was relatively greater in the previously less productive areas, where the knowledge gap is likely to have been the greatest. These findings were consistent with the contingent valuation findings: a vast majority of farmers, among both the current recipients and nonrecipients, were willing to pay for advice, indicating an unmet demand. However, the perceived value of the service, in terms of the amount offered, was well below what the government was spending on delivering it.

C.5.5 Policy implications

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

To raise program efficiency, a leaner and less intense presence of extension agents with wider coverage is likely to be required. There are not enough technical innovations to warrant a high frequency of visits, and there is unmet demand from those currently not receiving services. The program's blanket approach to service delivery, relying predominantly on a single methodology (farm visits) to deliver standard simple messages, also limits program efficiency. Radio programs are now popular, younger farmers are more educated, and alternative providers (non-governmental organizations) are beginning to emerge in rural Kenya. A flexible pluralistic approach to service delivery, particularly one that uses lower-cost means of communication, is likely to enhance the cost effectiveness of the program.

Finally, the main findings pointed to the need for institutional reform. The central focus of the institution should be the farmer. Decentralization of program design, including participatory mechanisms that give voice to the farmer (such as cost sharing and farmer organizations) should become an integral part of the delivery mechanism. Financial sustainability is critical. The size and intensity of the service

should be based on existing technological and knowledge gaps and the pace of flow of new technology. Cost recovery, even if only partial, offers several advantages: it provides appropriate incentives, addresses issues of accountability and quality control, makes the service more demand-driven and responsive, and provides some budgetary respite. Such decentralized institutional arrangements remain unexplored in Kenya and in many extension programs in Africa and around the world.

C.5.6 Evaluation costs and administration

Costs. The total cost of the evaluation was approximately $350,000, 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 approximately $160,000 for World Bank staff time and travel costs for data processing, analysis, and report writing should be added to reflect fully 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.

C.5.7 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 preproject stage and also independent sources of data for comparative purposes.

Cross-validation of conclusions using different analytical approaches and data sources is important to gather a credible body of evidence. Imperfect data and implementation problems place limits on the degree of confidence with which individual methods can provide answers to key evaluative questions. Qualitative and quantitative assessments strongly complement each other. The experience from this evaluation indicates that even in the absence of participatory beneficiary assessments, appropriately designed questions can be included in a survey to collect qualitative as well as quantitative information. Such information can provide useful insights to complement quantitative assessments.

If properly applied, contingent valuation can be a useful tool, especially in evaluating the value of an existing public service. The results of the application in this evaluation are encouraging, and the responses appear to be rational and reasonable.

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.

Case Study C.6 Evaluating Nicaragua's School Reform: A Combined Quantitative-Qualitative Approach

C.6.1 Introduction

In 1991, the Nicaraguan government introduced a sweeping reform of its public education system: it decentralized school management (decisions on personnel, budgets, curriculum, and pedagogy) and transferred financing responsibilities to the local level. Reforms were phased in over time, beginning with a 1991 decree that established community-parent councils in all public schools. Then, a 1993 pilot program in 20 hand-picked secondary schools transformed these councils into school management boards with greater responsibility for personnel, budgets, curriculum, and pedagogy. By 1995, school management boards were operational in 100 secondary schools and more than 300 primary schools, which entered the program through a self-selection process involving a petition from teachers and school directors.

As school management becomes more democratic and participatory and locally generated revenues increase, spending patterns were expected to become more supportive of efforts that directly improve pedagogy and boost student achievement.

The evaluation of the Nicaraguan school autonomy reforms represents one of the first systematic efforts to evaluate the impact of 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 decisionmaking processes. The evaluation also illustrates best practice techniques when there are no baseline data, and when selective (nonrandom) application of reforms rules out an experimental evaluation design.

The purpose of the qualitative component of the evaluation was to determine whether or not the intended management and financing reforms were actually observed in schools, and to assess how various stakeholders viewed the reform process. The quantitative component addressed the following question: Did changes in school management and financing actually produce better learning outcomes for children? The qualitative results showed that successful implementation of the reforms depended largely on school context and environment (that is, the poverty level of the community), while the quantitative results suggested that increased decisionmaking by schools was in fact significantly associated with improved student performance.

C.6.2 Evaluation design

The design of the Nicaraguan school autonomy reform evaluation is based on matched comparison, where data for a representative sample of schools participating in the reform process are compared with data from a sample of nonparticipating schools. The sample of nonparticipating schools is chosen to match as closely as possible the characteristics of the participating schools and hence provides the counterfactual. This design was chosen because the lack of baseline data ruled out a before-and-after evaluation and because reforms were not applied randomly to schools, thus ruling out an experimental evaluation design.

C.6.3 Data collection and analysis techniques

The qualitative study draws on data for a sample of 12 schools, 9 reformers, and 3 nonreformers that represent the control group.[11] The sample of 12 schools was picked to represent both primary and secondary schools, rural and urban schools, and schools with differing degrees of actual autonomy in decisionmaking. A total of 82 interview and focus group sessions were conducted, focusing on assessing how school directors, council members, parents, and teachers understood and viewed the decentralization process. All interviews were conducted by Nicaraguans, trained through interview simulation and pilot tests to use a series of guided questions without cueing responses. Interviews were audio-recorded, transcribed, and then distilled into a two- to four-page transcript that 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 (November-December 1995, and April-August 1997), and student achievement tests for students in these schools, which were conducted in November 1996. The school survey collected data on school enrollment; repetition and dropout rates; physical and human resources; school decisionmaking; and characteristics of school directors, teachers, students, and their families. The school decisionmaking module is unique, presenting a series of 25 questions designed to gauge whether and how the reform had actually increased decisionmaking by schools. The survey covered 116 secondary schools (73 reformers and 43 nonreformers representing the control group), and 126 primary schools (80 reformers and 46 nonreformers). Again, the control groups were selected to match the characteristics of the reform schools. The survey also gathered data for 400 teachers, 182 council members, and 3,000 students and their parents, with 10 to 15 students chosen at random from each school.

Quantitative data analysis used regression techniques to estimate an education production function. It examined the impact of the degree of school autonomy on student achievement levels, controlling for school inputs and household and student characteristics. The analysis measured the effect of both de jure and de facto autonomy. De jure autonomy was defined simply as whether or not the school has legally joined the reform, while de facto autonomy measured the degree of actual autonomy achieved by the school as the percentage of 25 key decisions made by the school itself. De facto autonomy was expected to

vary across schools because reforms were phased in (so schools in the sample were at different stages in the reform process), and because the capacity to implement reforms successfully varied according to school context, as found in the qualitative study).

C.6.4 Results

The qualitative study highlighted that policy changes at the central level did not always result in changes at the local level. In general, reforms were associated with increased parent participation, as well as with management and leadership improvements. But the degree of success with which reforms were implemented varied with school context. Of particular importance were the degree of impoverishment of the surrounding community—in poor communities, raising local school financing was difficult—and the degree of cohesion among school staff—where key actors such as teachers did not feel integrated into the reform process, decentralization was limited. Policymakers often ignore the highly variable local contexts into which new programs are introduced. The Nicaraguan results pointed out that the goal of increased local financing for schools was unlikely to be met in poor communities.

The quantitative study reinforced the finding that reform schools were indeed making more of their own decisions, particularly with regard to pedagogical and personnel matters. De jure autonomy—whether a school had signed the reform contract—did not necessarily translate into greater school level decisionmaking, nor did it affect schools equally. The degree of autonomy achieved depended on the poverty level of the community and on how long the school had been participating in the reform process. The regression results showed that de jure autonomy had little bearing on student achievement outcomes, but de facto autonomy was significantly associated with improved student achievement. Furthermore, simulations indicate that increased school autonomy had a stronger bearing on student achievement than other typical actions, such as increasing the number of textbooks, expanding teacher training, and reducing class size.

C.6.5 Policy implications

The evaluation results provided concrete evidence that Nicaragua's school reform produced tangible results. Reform schools indeed made more decisions locally and enhanced local decisionmaking resulted in improved student achievement.

The results also pointed out areas where policy can be improved, and the Ministry of Education introduced a number of changes in the school reform program. The program now places greater emphasis on the training of teachers and on promoting the pedagogical aspects of the reform. Further, in response to the financing problems of poor communities, the Ministry developed a poverty map–driven subsidy scheme. Finally, the tangible benefits from this evaluation prompted the Ministry to incorporate a permanent evaluation component into the reform program.

C.6.6 Evaluation costs and administration

Costs. The total cost of the evaluation was approximately $495,000, representing less than 1.5 percent of the World Bank credit that supported the reforms.[12] Of this total evaluation cost, 39 percent was spent on technical support provided by outside consultants, 35 percent on data collection, 18 percent on World Bank staff time, and 8 percent on travel.

Administration. The evaluation was carried out jointly by the Nicaraguan Ministry of Education, the World Bank, and researchers from the Harvard School of Education.

C.6.7 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 representative 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 achieved while others had not, and hence helped guide policy adjustments. Furthermore, because it was more intuitive, the qualitative work was more accessible and therefore captured the attention of Ministry staff, which in turn facilitated rapid capacity building and raised the credibility of the evaluation process within the Ministry.

Importance of Local Capacity Building. Local capacity building was costly and required frequent contact and coordination between Nicaraguan staff, 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 positive or negative. These evaluation results provided direct input into the reform as it evolved. The policy impact of the evaluation was also enhanced by a cohesive local team in which evaluators and policymakers worked collaboratively and by the support of the Minister of Education.

Sources and further reading:

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.

———. 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. 2000. "Evaluating Nicaragua's School-Based Management Reform." In Michael Bamberger, ed., *Integrating Quantitative and Qualitative Methods in Development Research.* Washington, D.C.: World Bank.

Case Study C.7 Schooling Outcomes in Philippine Elementary Schools: Evaluation of the Impact of Four Experiments

C.7.1 Introduction

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

The evaluation is noteworthy in that it explicitly aimed to build capacity in the host country so that evaluation would become an integral component of new initiatives, and data requirements would be considered before rather than after project implementation. Another major contribution of the evaluation was to check for robustness of results with different econometric approaches. Finally, the benefit-cost analysis applied at the end was important in that it explicitly recognized that significant results do not suffice to justify an intervention: inexpensive interventions may still be better than expensive ones.

C.7.2 Evaluation design

The key objective of the research was to evaluate the impact of the four different interventions on dropout rates and student outcomes. The evaluation design was conditioned by pragmatic as well as programmatic needs. The DIP team followed a three-stage school selection process:

- Two low-income districts were identified in each of five regions of the country.
- In each district, the team selected three schools which a) had all grades of instruction, with one class per grade; b) had a high dropout rate; and c) had no school feeding program in place.

- In one district, the options available were MLM, MLM-PTP, or nothing; in the other, SL, SL-PTP, or nothing.

The three schools in each district were assigned to the control group or to one of the two interventions based on a random drawing.

Pre-tests and post-tests in three subjects (mathematics, Filipino, and English) were administered at the beginning and end of the 1991 and 1992 school years to all classes in all 30 schools.

C.7.3 Data collection and analysis techniques

Baseline data collection began in 1990–91, and the interventions were implemented in 1991–92. Detailed information was gathered on 29 schools, on some 180 teachers, and on about 4,000 pupils in each of the two years. Although the questionnaires were very detailed, much of the detail turned out to be unnecessary: only a small subset of the information was actually used, suggesting that the burden of the evaluation process could be reduced.

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

The structure of the sampling procedure raised some interesting econometric problems for dropout rates and 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 controls in different ways.

In the analysis of dropout rates, it is natural to use 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 the result of small sample sizes. This is not uncommon with this type of procedure, and it is partly due to 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 and family background. The main result was unchanged. However, the regression technique did uncover an important indirect core cause of dropping out, which was poor academic performance. This naturally led to the second set of analysis, which focused on achievement.

Several econometric issues arose in the evaluation of the impact of the intervention on the academic performance of an individual in a given school at a given time: accounting for the clustered correlation in errors that is likely to exist for students in the same classes and schools; capturing unobserved heterogeneity; and dealing with selection bias (if students with lower academic performance are more likely to drop out of school, then estimates of program effects may be biased upwards).

The first issue was dealt with by applying a Huber-White correction to the standard errors. The second issue could in principle be dealt with at the individual level by using the difference in test scores as an independent variable. However, the authors argued that this was inappropriate because it presupposed that the value of the coefficient on academic performance was 1, which was not validated by tests. They therefore used academic performance in the first period as an explanatory variable, but this raised the problem of correlation with the error term, or endogenous regressor bias. This was handled by using as an instrumental variable for the pre-test score in each subject the pre-test scores in the other subjects. The authors noted, however, that the reduction in bias came at the cost of a reduction in efficiency, and hence reported both least squares and instrumental variables results. The authors used both school and teacher fixed effects to control for unobserved heterogeneity in learning environment and classroom conditions.

The third problem, selection bias, is one that is also commonly found in the literature, and for which there is no fully accepted solution. Clearly, 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 (Heckman 1976, 1979), 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 on an appropriate alternative. One solution proposed by Krueger (1997) is to impute test scores for students who exit the sample by assigning to them in years when they were absent from the sample their most recent test percentile. The authors report three sets of results: the simple regression of outcomes against intervention, the Krueger approach, the Heckman procedure, and instrumental variables results obtained using Krueger's procedure.

C.7.4 Results

The study found that the positive effect of multilevel learning materials—particularly with a parent-teacher partnership—on dropout rates and academic performance was robust to different specifications. The effect of school lunches was, in general, weak. An interesting component of the study was a cost-benefit analysis, that made the important point that significant results were not the whole story. A straightforward calculation of both the direct and the indirect (opportunity) costs of the programs led to the conclusion that the MLM approach was 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—targeting, for example, only malnourished or under-privileged children—might be more cost-effective. Furthermore, since the period of time between the implementation and the evaluation of the program was quite short, the evaluation could not address the long-term impact of the interventions.

C.7.5 Administration

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.

C.7.6 Lessons learned

Several lessons were learned through this evaluation. A major one was that a lot of vital longitudinal information can be lost if the identifiers of observations (in this case, pupils) are not unique over time, is lost. A second lesson was that very little of the information that was gathered in detailed surveys was used and that the burden on the respondents could have been substantially reduced. Third, the study highlighted the value of different econometric approaches and the advantages of finding consistent results across techniques. Fourth, the study was exemplary in both identifying and valuing the costs of the different interventions. Finally, although errors were clearly made during the study, the authors noted that a prime objective was to build evaluation capacity in the Philippines.

Sources and further readings:

Heckman, James J. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models." *Annals of Economic and Social Measurement* 5:475-492.

Heckman, James J. 1979. "Sample Selection as a Specification Error." *Econometrica* 47:153-161.

Krueger, Alan. 1997. "Experimental Estimates of Educational Production Function," NBER Working Paper 6051, Cambridge, MA

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):493-508.

Notes

1. An earlier version of this case study was prepared by Margaret Kakande, Head, Poverty Monitoring Unit, Ministry of Finance, Planning and Economic Development, Government of Uganda and Kimberley McClean, Head International Projects, Aus Health International.

2. For more information on the National Service Delivery Survey, see the World Bank Research Department Web site on public service delivery at http://www.worldbank.org/research/projects/publicspending/tools/tools.htm#Quantitative Service Delivery.
3. See http://www.uppap.or.ug/.
4. Available at http://www.worldbank.org/afr/findings/english/find142.htm.
5. The EDS survey was financed under a World Bank-supported project different from Trabajar. 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.
6. The EDS questionnaire is very comprehensive, collecting detailed data on household characteristics that help predict program participation and facilitate the use of the propensity scoring technique. The propensity score is calculated for each observation in the participant and control group sample using standard logit models.
7. This is a composite index that includes residential crowding, sanitation facilities, housing quality, educational attainment of adults, school enrollment of children, employment, and dependency (ratio of working to nonworking family members). The index was somewhat dated, although this had the advantage that the departmental poverty measure was not influenced by Trabajar interventions.
8. No attempt was 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 nonuse values for a variety of public and private goods and services. Interviewees are asked to state their willingness to pay to avoid a hypothetical change in the provision of the goods or services, that is, the "contingent" outcome. In this case, farmers were asked how much they would be willing to pay for continued agricultural extension services, should the government cease to provide them.
10. These three surveys generate a panel dataset for approximately 300 households. The surveys cover household demographics, farm characteristics, and input-output data on agricultural production; the 1990 and 1998 surveys also collected information on contact with extension services, including awareness and adoption of extension messages.
11. Data were actually gathered for 18 schools, but only 12 of these schools were included in the qualitative analysis because of delays in getting the transcripts prepared, and a decision was made to concentrate the bulk of the analysis on reform schools, which provided more relevant material for the analysis.
12. This total does not include the cost of local counterpart teams in the Nicaraguan Ministry of Education.