Building Educational Evaluation Capacity in Developing Countries

John Middleton,
James Terry,
and Deborah Bloch

Education monitoring, testing, and evaluation systems are urgently needed in developing countries. Lessons from the American experience suggest that investment plans should tie evaluation standards, requirements, and funding to program support for policy and institutional changes at the national, intermediate and school levels of education systems.
Relatively few developing countries have established a sustainable capacity for educational monitoring and evaluation. As a result, the efforts of governments and assistance agencies to improve education have been hampered by lack of information on outcomes and costs.

World Bank experience with educational evaluation has been disappointing, emphasizing monitoring of project inputs with little attention to outcomes and costs. Efforts to develop a sustained evaluation capacity have largely failed.

In the United States, educational evaluation has become an integral part of education management. Large-scale research studies have given way to student achievement testing and localized evaluations that provide decision-makers with information useful in improving the quality of schools.

These results were achieved through federal policy and funding support for evaluation. Standards, requirements, and financial incentives for evaluation were tied to federal financing for school improvement programs. Institutional capacity as developed at state, district, and school levels.

Similar investment strategies can be used in developing countries. However, the strategies should be incremental, first putting systems in place at relatively low levels of technical sophistication, then raising the technical level as institutional capacity is developed. Staff training and development, including the social science and education faculties of universities, is essential to developing sustainable capacity.
## Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>INTRODUCTION</strong></td>
<td>1</td>
</tr>
<tr>
<td>Definitions</td>
<td>2</td>
</tr>
<tr>
<td><strong>THE EXPERIENCE OF THE WORLD BANK</strong></td>
<td>4</td>
</tr>
<tr>
<td>Overview of M&amp;E Performance</td>
<td>4</td>
</tr>
<tr>
<td>Monitoring and Evaluation in Projects: Design and Achievements</td>
<td>5</td>
</tr>
<tr>
<td>Recent and Past Practice in M&amp;E Design</td>
<td>5</td>
</tr>
<tr>
<td>Achievements</td>
<td>7</td>
</tr>
<tr>
<td>Summary</td>
<td>12</td>
</tr>
<tr>
<td><strong>THE U.S. EXPERIENCE</strong></td>
<td>12</td>
</tr>
<tr>
<td>Origins of Educational Evaluation in the United States</td>
<td>13</td>
</tr>
<tr>
<td>School Accreditation</td>
<td>14</td>
</tr>
<tr>
<td>Impact of Federal Legislation</td>
<td>15</td>
</tr>
<tr>
<td>Limitations in Available Methodologies</td>
<td>16</td>
</tr>
<tr>
<td>Forces for Change</td>
<td>17</td>
</tr>
<tr>
<td>Improved Utilization</td>
<td>17</td>
</tr>
<tr>
<td>New Methodologies</td>
<td>19</td>
</tr>
<tr>
<td>Changes in Legislation and Funding:</td>
<td>24</td>
</tr>
<tr>
<td>The Rise of Local Accountability</td>
<td>24</td>
</tr>
<tr>
<td>Summary</td>
<td>27</td>
</tr>
<tr>
<td><strong>LESSONS FOR THE IMPROVEMENT OF MONITORING AND EVALUATION</strong></td>
<td>29</td>
</tr>
<tr>
<td><strong>IN DEVELOPING COUNTRIES</strong></td>
<td>29</td>
</tr>
<tr>
<td>Strategies for Building M&amp;E Capacity</td>
<td>29</td>
</tr>
<tr>
<td>Who Should be Served?</td>
<td>30</td>
</tr>
<tr>
<td>For What Purposes?</td>
<td>31</td>
</tr>
<tr>
<td>What Information is Needed?</td>
<td>32</td>
</tr>
<tr>
<td>What Kinds of Systems Are Needed?</td>
<td>33</td>
</tr>
<tr>
<td>How Can Monitoring and Evaluation Systems Be Established and Sustained?</td>
<td>34</td>
</tr>
<tr>
<td>A Conceptual Framework for Institutional Development</td>
<td>35</td>
</tr>
<tr>
<td>The U.S. Experience Interpreted</td>
<td>36</td>
</tr>
<tr>
<td>Implications for the Design of M&amp;E Systems</td>
<td>37</td>
</tr>
<tr>
<td><strong>CONCLUSIONS: OPERATIONAL ISSUES FOR THE WORLD BANK</strong></td>
<td>40</td>
</tr>
<tr>
<td><strong>BIBLIOGRAPHY</strong></td>
<td>42</td>
</tr>
</tbody>
</table>
EXECUTIVE SUMMARY

Despite more than two decades of financial and technical assistance, relatively few developing countries have established sustainable capacity for educational monitoring and evaluation. In consequence, the efforts of both governments and assistance agencies to improve education have been significantly weakened by the absence of information, notably on educational outcomes and on costs. The effects of investments are difficult to gauge; lessons of experience difficult to accumulate.

Directions for improved strategies for support of educational monitoring and evaluation are sought from two sources. One is the experience of the World Bank. The second is the history of the development of educational evaluation in the United States, where theory and practice are best developed and documented among developed nations.

The experience of the Bank has been disappointing, emphasizing monitoring of the delivery of project inputs, with very little attention to the outcomes or costs of education. Efforts to develop sustained institutional capacity for monitoring and evaluation have rarely succeeded. As strategies for investment in education become more complex, the inability of governments and, consequently, assistance agencies to monitor and evaluate the outcomes and costs of investments is an increasingly significant barrier to efficient use of scarce resources.

In contrast, over the past three decades monitoring and evaluation have become integral parts of the U.S. education enterprise. This development has been marked by the heavy influence of federal policy and funding, and by significant shifts in the nature and purpose of evaluation activity. The system that has emerged eschews large scale research evaluation in favor of localized evaluations built around achievement testing, and designed to provide decision-makers with a variety of information useful in improving education at the school and district level. Monitoring and evaluation is designed with a clear focus on the utilization of results, and has thus become an integral part of education management. It is a central element of current efforts to improve the quality of schools.

The paper concludes that while project monitoring will continue to be important, priority should be given to the establishment of sustained government capacity for monitoring and evaluation of the education system as a whole. This will require selective investment in capacity building in countries with strong administrative frameworks and a clear policy commitment to monitoring and evaluation. Priority should be given to establishing the capacity for achievement testing, to primary education and to sectoral operations.

Lessons from the U.S. experience show the importance of achievement testing, and of policies and incentives that tie standards, evaluation requirements and financial support for monitoring and evaluation to program support for educational change and improvement. Explicit institution building strategies are also necessary. These should highlight the evaluation function through institutional structures; build capacity at central, intermediate and school levels; build evaluation functions into job descriptions and incentive
schemes; support the professionalization of the evaluation field; and develop both the technical and managerial competence of staff. Strengthening the social science and education faculties of universities is needed to provide a pool of available technical expertise.

Institutional development strategies should be incremental. At the onset, all elements of a monitoring and evaluation system should be put in place for a significant sub-sector (or geographic region) of the education system. But the technical sophistication of these elements should be kept low. This will provide a core around which institutional development activities can be organized. As this basic system becomes effective, it can be extended in scope and sophistication.

These investments will require a higher level of analysis and design work than that normally made available through donor or government resources. The costs of such investments in M&E capacity will be significant. For the World Bank to be effective in addressing the absence of sustained monitoring and evaluation, renewed policy and managerial support will be required.
INTRODUCTION

Despite more than two decades of financial and technical support from bi-lateral and multi-lateral assistance agencies, relatively few developing countries have established sustainable capacity for educational monitoring and evaluation (M&E). Recent success in Kenya in developing and using evaluation information for educational improvement (Somerset, 1987), not only highlights the usefulness of monitoring and evaluation in improving the efficiency and effectiveness of education, but also raises once again the question of why M&E has not yet become widespread and effective tool for educational improvement in developing countries.

Nor, in the experience of the World Bank, has M&E been an effective tool for the improvement of development assistance operations. Weak monitoring of the "software" aspects of projects, combined with nearly universal absence of information on either educational or economic outcomes, has restricted the Bank's ability to learn systematically from experience.

In an era of constrained resources for education development, the lack of effective monitoring and evaluation significantly impairs efforts of governments and assistance agencies to make most effective use of the resources that are available in pursuit of the twin goals of improved quality and expansion of education systems. As investments in more complex strategies for educational improvement increase, including sector lending, learning from experience becomes even more important -- with a correspondingly more significant place for effective monitoring and evaluation.

This paper seeks directions for improved strategies for support of monitoring and evaluation from two sources. One is the experience of the World Bank. The second is the history of the development of educational evaluation in the United States, where theory and practice are perhaps best developed and documented among developed nations. The focus of the review is on policies and strategies for investment design; the technical details of evaluation design and implementation are treated only insofar as they affect broader policies and strategies.

The experience of the World Bank has been largely disappointing. Monitoring and evaluation in Bank education investments has focused on the delivery of project inputs; relatively little attention has been given to the assessment of educational outcomes. Investments in the development of sustained institutional capacity for system monitoring and evaluation have most often failed. As a result, governments lack the capacity for ongoing
monitoring and evaluation, and the capacity of the Bank to learn from its own project evaluation system has been weakened.

In contrast, over the past three decades evaluation has grown to become an integral part of the U.S. education enterprise. This period was marked by the heavy influence of federal policy and funding on a decentralized education system, and by significant shifts in the purposes and nature of evaluation activities. While not free of political or technical problems, the current practice of educational evaluation provides administrators a steady flow of information useful in improving the performance of schools.

The analysis is presented in four parts. The first examines monitoring and evaluation in past and present World Bank education investments. The U.S. experience is reviewed in the second. In the third lessons from the U.S. experience are developed and assessed against the realities of developing country contexts. A concluding section identifies operational implications of the analysis for the World Bank and other assistance agencies.

Definitions

Educational evaluation encompasses a variety of purposes, models and techniques.1 A major distinction is drawn between summative and formative evaluation. The purpose of the former is to develop information for judgments regarding the overall impact and effectiveness of programs, and thus on their continuation or replication. The purpose of the latter is to develop information during program implementation for redesign and improvement. Both terms are used most frequently in categorizing types of evaluation for interventions and innovations designed to change and improve educational practice.

A second distinction can be drawn between evaluation of specific interventions designed to change and improve education practice (often called "projects"), and ongoing evaluation activities designed to produce a regular flow of information on the system and its components. Project evaluation can be carried out as a free-standing effort where there is no ongoing, system-wide evaluation. Under these circumstances, however, the designs must provide substitute mechanisms needed to collect data on non-project schools, increasing complexity and cost. Scattered project evaluations seldom result in reliable cumulative findings that help guide improvement of the larger system.

Numerous models, or methodologies, for educational evaluation have been developed. These provide integrated frameworks for the conduct of a range of monitoring and evaluation activities. Many tie a comprehensive set of evaluation activities explicitly to both summative and formative decision-making. Other models include systems analysis; professional review (accreditation); the behavioral objectives or goals-based approach, in which

---

1 The intent here is to establish a relatively few basic definitions in a complex field. A full review may be found in House, 1980.
achievement of specified objectives is assessed; and the contrasting goal-free approach, in which all effects of a program are assessed, whether goal-related or not.

Models of evaluation are often confused with techniques for design, data collection and analysis. In general, most techniques can be used within a given model. Research evaluation, or evaluation research, is a term most closely identified with the use of the techniques of experimental and quasi-experimental designs, pre-post measures and multi-variate statistical analysis for summative purposes. Monitoring, inseparably with the aid of management information systems is comprised of a set of techniques (indicators, surveys, routine reporting, observation) that generate information for formative purposes, as well as a data base useful, in combination with other types of evaluation, for summative judgments.

Student achievement testing and examination systems provide crucially important information for evaluation of all kinds. If implemented routinely and for the education system as a whole, testing systems provide the information that enables administrators, teachers, parents and students to assess educational performance across time, and to compare the performance of different elements of the education systems -- schools, districts and states. Achievement data also provide a framework within which the effectiveness of specific interventions (such as new curricula) can be judged against past practice.

There is a tendency in the Bank, and elsewhere, to combine the terms monitoring and evaluation to form the hybrid "monitoring and evaluation" (M&E). These can be separate activities, each with its own purpose(s), models and techniques. Monitoring alone is essentially a formative management process, providing a flow of information for improved implementation -- including program re-design. Evaluation alone is summative, concerned principally with outcomes and less with processes. Ideally, the two are linked together in an integrated design which provides information on both processes and outcomes, strengthening the ability of decision makers to not only assess outcomes, but trace the causal relationships between programs as actually implemented and achievements.
The World Bank and its Borrowers have faced two distinct but interrelated evaluation tasks. The first is the monitoring and evaluation of specific investment projects. The second is the development of the capacity of Borrower agencies for sustainable monitoring and evaluation of the performance of the education system and its components. In some cases, investments in project monitoring and evaluation have been intended to contribute to sustained institutional capacity.

Overview of M&E Performance

Since 1979, annual reviews of project evaluations conducted by the Bank's Operations Evaluation Department (OED) have indicated that while monitoring of the delivery of physical inputs has been reasonably effective, the same has not been true for the "softer" aspects of projects -- curriculum development and effectiveness, teacher training, and institutional development. Evaluation of educational outcomes has been largely lacking. And the reviews find little impact on sustainable Borrower evaluation capacity.

The continuing lack of Borrower capacity for monitoring and evaluation in turn significantly weakens the development of an information base for specific projects, with the result that project completion evaluations suffer from inadequate data, most particularly on unit costs, educational outcomes and software component performance. These weaknesses in project evaluations have hampered the operational reviews through which the Bank draws lessons from past experience to improve lending operations. (Johanson, 1986; Romain, 1986; Romain and Armstrong, 1987; Middleton and Demsky, 1988).

In an earlier era, when World Bank projects primarily financed civil works and equipment, weak Borrower capacity for monitoring and evaluation was, perhaps, less troublesome in its impact on sustainable development. However, since the early 1970s Bank investments in education have increasingly supported system change and improvement through curriculum development, teacher training and management and organizational development. In a study of these change components, Verspoor found that the most successful investments focused "... on the institutionalization of a capacity to manage change and the development of systems and structures capable of dealing effectively with unexpected events." (1988, forthcoming, p. ix). A steady flow of information on performance of the project at all levels, most especially the classroom, was important to success. The study concludes that one of the four essential characteristics of effective educational change programs -- and thus of institutionalization of effectiveness and efficiency measures -- is the establishment of effective mechanisms for the feedback of information on performance to support enroute program modification and improvement. Among such mechanisms, those that support regular collection and analysis of the outcomes of education -- student achievement, employment -- are centrally important.
Yet evaluation of outcomes has been weakly supported. Only 08% of the project components studied by Verspoor supported the development of testing systems. In a review of 76 World Bank investments in vocational education and training, Middleton and Densky (1988) found the same low level of attention to examination and certification processes, and to assessment of the employment outcomes of training. In general, this lack of attention to assessing educational and economic outcomes is a clear indicator of low levels of investment in the development of permanent evaluation capacity, by Borrowers and the Bank alike.

Monitoring and Evaluation in Projects: Design and Achievements

To gain more specific insight into the design of Bank support for monitoring and evaluation (both for projects and for broader system capacity), a sample of ongoing and completed projects was studied. Current design practice was identified through a review of the 16 project appraisal reports approved in FY85-86. The design and effectiveness of past investments were assessed through an analysis of the appraisal reports and project evaluations for the 17 education sector projects completed in the same period.

The review of investment design in completed and ongoing projects indicates increasing attention to the design of project monitoring systems at the appraisal stage. Evaluation of outcomes, however, has received very little attention. The data suggest that efforts to build Borrower capacity for system-wide M&E may be declining.

The record of success in completed projects has been very weak for project monitoring and evaluation as well as for the development of system capacity. Reasons include inappropriate designs, and lack of the policy and incentive support that might overcome Borrower reluctance to support monitoring and evaluation efforts.

Recent and Past Practice in M&E Design

This component of the analysis focused on the design of support both for project monitoring and evaluation and for the development of system capacity in both ongoing and completed projects. The basic question was whether or not provision was made for project monitoring and evaluation, development of M&E capacity, or both. Additional questions regarding investment design were:

a) Were indicators specified?
b) Was responsibility for M&E clearly assigned?
c) Were procedures for administration of M&E specified?
d) Were decision points for use of M&E results specified?
e) Were policy support and/or incentives provided?

2 A list of the projects reviewed may be found in Annex I.
The nature of the inputs designed to support M&E and the nature of M&E objectives and methodology were also reviewed.

Although Bank appraisal reports provide relatively little information, the design of investments in system capacity for M&E appears to have taken three principal forms. One is as part of overall efforts to develop institutional capacity for educational planning. Support for M&E is usually included. Second has been support for monitoring and evaluation units in line ministries. The third has been the inclusion of M&E responsibilities in specialized education support institutions, such as curriculum development centers.

The inputs to these efforts have been specialist services, fellowships and equipment (notably computers). The same inputs have been used to support project M&E. In addition, special studies were included at appraisal in about half of the 33 projects. Rationales for any particular combination of inputs have been weak and inconsistent.

The data on specific aspects of design indicate increased provision of project monitoring, and somewhat less attention to building system capacity in the later projects (see Table 1). Evaluation of learning or economic outcomes has received little attention. On the surface, at least, more attention has been given in recent project appraisals to planning for monitoring, as seen in the assignment of responsibility, specification of administrative procedures and decision points, and the increased use of indicators.

In the projects studied, the net effect of assignment of responsibility and specification of procedures has been to design for centralized monitoring systems that provide project information needed by the Borrower and the Bank. Decision points specified are invariably points during project implementation at which the Bank and the Borrower will consult on implementation progress. At least as reflected in project appraisals, other levels of the education system -- states, districts, schools -- figure in the process only as sources of information.

The increase in the specification of progress indicators as the basis for monitoring at appraisal is encouraging, indicating stronger levels of discussion between the Bank and Borrowers on the nature of the monitoring and evaluation effort. At the same time, the indicators used demonstrate convincingly that the Bank continues to ignore factors crucial to project evaluation. Of the 100 indicators used in the FY85-86 appraisals, there was one for student achievement, one for completion of a tracer study, and one for unit recurrent costs. The remaining 97 were about equally divided between the provision of inputs and the accomplishment of output objectives, most often enrollments.

Indicators thus support monitoring of project implementation progress, but not evaluation of project impact or costs.
Table 1: Comparison of Monitoring and Evaluation Design in Appraisal Reports for Two Time Periods

<table>
<thead>
<tr>
<th></th>
<th>Completed in FY85-86</th>
<th>Approved in FY85-86</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Projects</td>
<td>17</td>
<td>16</td>
</tr>
<tr>
<td>Project M&amp;E Supported</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td>M&amp;E System Development</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>Indicators Provided</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>Responsibility for M&amp;E Assigned</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td>Administrative Procedures Specified</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Decision Points Specified</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Policy or Incentive Support Provided</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Relatively low levels of attention to evaluation are also reflected in the limited discussion of M&E objectives and methodologies found in the appraisal reports of both eras. In general, evaluation of impact is restricted to a few special studies, most to be carried out by consultants.

With the exception of three sector loans (in Korea, Malaysia and Morocco), policy and incentive provisions designed to support the development of monitoring and evaluation have been absent from Bank investments. In these three cases criteria for sub-project evaluation and monitoring formed a policy-level context; and, in the case of Morocco, release of the tranche of the loan is conditioned on an interim project review, providing some incentives for project evaluation.

Achievements

Of the seventeen projects completed in FY85-86, twelve made provision for either project M&E or capacity development, or both. The fate of these investments, as shown in Table 2, was grim. All but two failed.
The two successful investments came in Lebanon and Korea. While information is thin in completion reports, a number of factors in success are apparent. Both nations had above average implementation capacity. In both cases, attention was focused on monitoring systems. In Lebanon, a simple input of data processing equipment was assimilated and used. In Korea, a complex program of monitoring that included tracer studies and cost assessment was made effective through the establishment of a strong institutional base in specialized curriculum and research agencies with financial support for operating costs from the Bank. It is also worth noting that the accreditation system, which involved site assessments to determine if institutions conformed to quality standards, was hampered to some extent by lack of incentives for institutional compliance.

The causes for failure are readily apparent: weak implementation capacity, not compensated for by strong designs; and low levels of commitment to M&E, reflected in overt resistance on the part of the Government and in reduction or elimination of TA resources. Some of this resistance may be attributable to skepticism regarding the utility of long and complex evaluation studies to the political reality of decision-making; some to the reluctance to be "evaluated" on the success of difficult and unpredictable projects and programs.

The low level of attention to evaluation of educational outcomes noted for the recent projects in the sample is replicated in the completed projects. On the Bank's side, lack of attention to evaluation of project outcomes may reflect the difficulty that has been encountered in attempting to go beyond monitoring of project inputs and processes. Recent case studies of three such efforts in Bank projects document serious problems in design and utilization (Searle, 1986). Among these were overly complex designs, lack of attention to unstable implementation processes, and failure to address the complex political and bureaucratic decision-making context in which evaluation information is (or is not) used.
<table>
<thead>
<tr>
<th>Country</th>
<th>Design</th>
<th>Inputs</th>
<th>Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afghanistan</td>
<td>Develop staff capacity; Evaluation system for project institutions</td>
<td>Studies, Fellowship Training, 10 MY TA</td>
<td>Project cancelled due to political conditions</td>
</tr>
<tr>
<td>Algeria</td>
<td>Continuous evaluation system; tracers</td>
<td>90 months TA for studies and system design</td>
<td>Major implementation problems; studies completed; system not developed</td>
</tr>
<tr>
<td>Congo</td>
<td>Reinforce educational planning</td>
<td>5 years TA; 14 years fellowships; vehicles &amp; equipment</td>
<td>Major implementation problems; TA reduced; evaluation not developed</td>
</tr>
<tr>
<td>Haiti</td>
<td>Studies; curriculum evaluation; government agrees to establish tracer system</td>
<td>3 years TA</td>
<td>Studies eliminated in favor of staff training; tracer system not established; MOE refuses to support evaluation objectives</td>
</tr>
<tr>
<td>Honduras</td>
<td>Establish system to evaluate project schools; to be expanded to national system</td>
<td>4 months TA</td>
<td>System not established because Ministry refused to hire consultant</td>
</tr>
<tr>
<td>Kenya</td>
<td>Evaluation system for craft centers; curriculum evaluation; computerize examinations</td>
<td>6 months TA; equipment; incremental operational costs</td>
<td>Curriculum evaluation established; craft center system not developed due reluctance to use TA and late equipment procurement; data processing not installed</td>
</tr>
<tr>
<td>Country</td>
<td>Design</td>
<td>Inputs</td>
<td>Outcomes</td>
</tr>
<tr>
<td>---------</td>
<td>--------</td>
<td>--------</td>
<td>----------</td>
</tr>
<tr>
<td>Korea</td>
<td>Establish permanent capacity for manpower monitoring, curriculum evaluation, accreditation of educational institutions. Criteria and procedures for evaluating sector sub-projects. Specific responsibilities assigned to 3 agencies, including one new one. Quarterly reporting.</td>
<td>Establishment &amp; initial operating costs of M&amp;E agencies; TA, studies</td>
<td>M&amp;E effective and used in project management. Accreditation system established, though lack of incentives for institutional acceptance and underfunding weakened the effort</td>
</tr>
<tr>
<td>Lebanon</td>
<td>Computerized MIS; tracers</td>
<td>Equipment</td>
<td>MIS installed</td>
</tr>
<tr>
<td>Madagascar</td>
<td>Build educational planning in decentralizing system, school inventory, data processing; evaluation Task Forces to organize formative evaluation</td>
<td>TA, fellowships</td>
<td>TA and fellowships eliminated because of cost over-runs; financed instead by UNICEF with UNESCO execution. Only one Task Force functioned. Impact on capacity slight. Data processing system dropped.</td>
</tr>
</tbody>
</table>
Table 2: Outcomes of Investments in Monitoring and Evaluation Systems, Projects Completed in FY85-86 (continued)

<table>
<thead>
<tr>
<th>Country</th>
<th>Design</th>
<th>Inputs</th>
<th>Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mali</td>
<td>Evaluation in each of the 5 components. All focused on project activities. National Pedagogical Institute for project curricula; AG training evaluation specialist; Enterprise training needs survey system by National Management Institute; Regular evaluations of literacy program; Follow-up of pre-investment study. National Research Team</td>
<td>5 Years TA</td>
<td>Almost ignored in PCR. Some studies conducted; no evidence of utilization</td>
</tr>
<tr>
<td>Senegal</td>
<td>Evaluation of all project activities. Build educational planning capacity. Case studies, &quot;indepth evaluation;&quot; tracers. Government and IDA to consult on findings</td>
<td>TA. Evaluation Specialist</td>
<td>Severe implementation problems. Evaluation not done as local staff not trained and found evaluation threatening. Tracer system not established.</td>
</tr>
<tr>
<td>Somalia</td>
<td>Build permanent evaluation system</td>
<td>2 Years TA - Evaluation Specialist; Fellowships; Local training</td>
<td>Local training provided; fellows fall to return to their posts.</td>
</tr>
</tbody>
</table>
Summary

While planning for monitoring of project inputs and processes has improved, it has generally been restricted to centralized information systems to support the project information needs of the Ministry and the Bank. Other levels of education have benefitted little. Evaluation of outcomes is largely lacking. Very little progress has been made in the establishment of sustainable evaluation capacity, in part due to weak investment designs, and in part to perceived lack of utility of evaluation on the part of Borrowers. The strong policy and incentive support for project M&E and the development of system capacity that might address such reluctance has not yet become part of lending strategies. The Bank's experience with more complex research evaluation studies has not been encouraging.

This situation gives rise to considerable concern in an era when educators in developing countries are challenged to improve system efficiency and effectiveness under conditions of scarce resources. And because meeting this challenge will require more complex strategies of educational change under conditions of great uncertainty (political, financial and technical), the need for effective evaluation is greater than in the past. Indeed, the development of testing and monitoring systems has been identified as a critical input to the improvement of education in Sub-Saharan Africa (World Bank, 1987).

Finally, the evolving nature of Bank lending for education must be taken into account. The growing emphasis on sector and policy lending shifts the focus of the dialogue between Borrowers and the Bank away from investments in specific institutions towards policy and organizational changes to improve system quality and efficiency. The details of implementation in these kinds of investments are increasingly left to Borrower determination. And while more traditional investment projects will continue to be an important part of the portfolio, resource constraints in Bank operations have reduced the amount of staff time available for preparation and supervision. In short, the ability of Bank staff to engage with Borrowers in the development of evaluation capacity is likely to be reduced. Under these circumstances, both Bank and Borrower staff need to have a clear sense of what kinds of evaluation capacity are most useful, and of the practical steps that might be taken to put it in place.

THE U.S. EXPERIENCE

The experience of the United States with educational evaluation is, in most respects, the longest and best documented among all developed countries, in no small part because of the relatively high levels of financial support that have been made available through federal sources. Indeed, much of current world practice in educational evaluation is based on work originating in the U.S. And while at first glance the resource-rich context of American educational evaluation may appear to have little relevance for developing countries, our review suggests that there are useful insights regarding the purposes and approaches to evaluation that education authorities
may consider in seeking to stimulate and sustain effective evaluation practice.

Much of this insight arises from the fact that the theory and practice of educational evaluation in the U.S. have changed considerably over the past twenty-five years, both in response to changes in the policy context for evaluation and as a result of dissatisfaction among the evaluation community with the approaches and methodologies with which the field began. The changes that have taken place have led to a clear focus on the generation of information useful for decision-making at the school and district level, and consequently, on achievement testing and on pragmatic and responsive evaluation designs. The evaluation practice that has emerged is thus oriented toward incremental improvements in education quality; large-scale research-evaluation studies seeking generalizable truths about education have largely disappeared. These changes bring current practice in the U.S. close to the kinds of evaluation that will be needed as developing countries turn their efforts to wards quality and efficiency through programs of educational change and improvement.

The role of the federal government in stimulating and developing evaluation practice has been crucial over the past twenty-five years. The structure of much of this support has been analogous to World Bank investment projects in that funds are supplied to States and localities to support interventions aimed at raising the quality of schooling. The history of this involvement yields insight into the possible effects of policy and financial incentives for evaluation. Importantly, as developing countries consider and in some cases initiate the decentralization of education management, lessons from the highly decentralized U.S. system become more relevant.

The discussion begins with an historical overview of the origins of educational evaluation in the U.S. Three major forces for changed practice are then discussed: the problem of utilization, the problem of methodologies, and the effects of changes in legislation and funding, including the rising importance of the accountability of schools for student performance.

Origins of Educational Evaluation in the United States

Educational evaluation in the United States developed in response to the social and political circumstances that defined the arena within which it matured. As one writer recently noted, evaluation, "unlike many other sciences.. makes no distinction between applied and basic research. The real world of social programs is its only home; there is no laboratory for evaluation practice." (Cook, 1986; p. 194).

Evaluation was an accepted part of the American educational scene before the turn of the century. Horace Mann directed an evaluation of the Boston schools in 1845 and Joseph Rice, between 1887 and 1898, conducted a study involving 33,000 students on the effects of the teaching of spelling. Standardized testing became popular in the early 1900s in response to a greater emphasis on the efficient use of educational resources and increasing interest in the accreditation of educational institutions.
The term "education evaluation" was coined by Ralph Tyler in the early 1930s. His writings on evaluation were instrumental in establishing a new approach to evaluation. He supported the goals of the Progressive Education movement of John Dewey and encouraged the pragmatic philosophy that set this group apart. His definition of evaluation called for a comparison of intended with actual educational outcomes for a given group of learners. This had the advantage of providing for internal assessment of the effectiveness of an education program in terms of its own goals, removing the need for costly comparisons between experimental and control groups, and foreshadowing current research on effective schools. Tyler's influence over the direction of evaluation was at its peak in the mid 1940s and continued for the next 25 years.

The decade between the late 1940s and 1950s saw little substantive activity within the field. Educational testing was expanded, and taxonomies of objectives and statistical procedures were developed, but relatively little effort was made to use these techniques to improve education. Methodologies focused on the measurement of program effects, using before and after measures, with very little attention to program processes. Federal and state governments were only marginally involved with education and evaluations, if done at all, were normally sponsored by local organizations.

School Accreditation

Although evaluation of the outcomes of schooling received little attention in this early period, considerable effort was devoted to the establishment and enforcement of standards for institutional quality. This emerged in the first fifteen years of the 20th century in the form of school accreditation (Young, et.al., 1983).

At the time, colleges faced a difficult problem: how to select among applicants from a large number of secondary schools with locally determined curricula, and with unknown levels of quality. The response was the development of curriculum and institutional quality standards, and a process through which individual secondary schools could be certified, or "accredited," on a voluntary basis. As the number of colleges participating in the system grew, it became incumbent on all secondary schools seeking success for their students in college admission to participate. Being "accredited" became the first mark of quality in the eyes of not only of college admissions officers, but of parents and students as well. The process became institutionalized on a regional basis, and was extended to include post-secondary institutions.

Accreditation continues as an important part of educational evaluation. It is managed by regional voluntary associations of schools and colleges. Officers of these associations are elected by member institutions, and are supported by a small professional staff. Each association sets standards of quality for both public and private institutions of different types -- colleges, vocational schools, specialized training institutions, and both private and public secondary schools. Standards are established for facilities, faculty qualifications and numbers, curricula, degree requirements, and financial resources. To maintain accredited status,
institutions must periodically (usually every 5 years) prepare a self-study that demonstrates achievement of standards and documents ongoing efforts for institutional improvement. This report is audited through a visit of external evaluators selected by the association. An independent audit report is considered by the association officers, who may vote to continue full accreditation, or issue one of a set of less comprehensive endorsements. These range from limited accreditation, with notice of specific defects and a time period for correction (one to three years), to a more serious warning, to removal of accredited status.

In recent years the focus of the accreditation process has become formative as associations seek to assist institutions in improving quality.

The process has perhaps been most effective for private schools and colleges, for which accredited status is important to attracting students. Accreditation has been useful in guarding the public against consistently inadequate institutions. Well-established universities and schools take the process less seriously, but in general cooperate. The voluntary, self-regulation model has proven attractive, enabling institutions to have a voice in the establishment of standards and in the evaluation process.

The accreditation model of institutional evaluation has also been adopted by a number of state education agencies as a complement to routine monitoring and other evaluation processes for public schools. In many of those cases it is extended to primary schools.

**Impact of Federal Legislation**

The technological challenge presented by the launching of Sputnik in 1957 resulted in the National Defense Education Act of 1958. This legislation significantly increased the role of the federal government in education as the sponsor of a variety of new science, mathematics and foreign language curriculum initiatives. Funding was also specifically provided for the evaluation of these efforts.

The Elementary and Secondary Education Act (ESEA) of 1965 further increased federal impact on evaluation. Of special importance was the requirement that specific evaluation provisions be included at each project site. This stipulation meant that in order to receive federal funds, each school district was required annually to submit an evaluation, using appropriate standardized test data, documenting the effectiveness of its local program in remediating the educational deficiencies of its disadvantaged students. In addition to mandating standardized test scores, the bill also called for an assessment of the congruence between educational outcomes and objectives. The effect was to "force evaluators out of the cool world of pure analysis into the maelstrom of political and administrative complexity." (Cook, 1986; p. 197).

The legislation also provided administrative funding to the State education agencies (SEAs) through which funding passed to school districts. This funding supported the establishment of permanent evaluation capacity in the SEAs in the late 1960s and early 1970s. These units were responsible for
ensuring that federal requirements for evaluation were met. Evaluation activities were conducted by the staff of these units and contracted out to consultants.

Importantly, the combination of available funding and localization of evaluation greatly expanded the opportunity for academics and education professionals to practice the craft of evaluation. Funding was a particular incentive to university faculty, who saw opportunities to apply social science research through the practice of educational evaluation. Evaluation developed as an academically respectable discipline, complete with professional associations and journals.

**Limitations in Available Methodologies**

The evaluation approaches available at the time, however, proved inadequate. Testing materials designed to accommodate and rank the academic abilities of the middle class student did a poor job of recognizing the ability or measuring the academic gains of academically and socially disadvantaged students. Evaluation designs were not able to distinguish or isolate the effects of projects, nor were they able to provide insight into the success or failure of implementation. Evaluators found themselves unable to provide evidence of the effects of the program treatment -- indeed, they were typically unable to identify the treatment or to specify the degree to which it had been implemented.

These shortcomings originated, in part at least, in the policy context of the era. The 1960s were a time of large social programs that rested on the belief that social science theory would point to the clear causes of target problems and suggest interventions (or projects) to overcome them. There was an implicit expectation that the various projects would be implemented and evaluated in a way that would provide unambiguous answers regarding effectiveness. It was further assumed that demonstrated and evaluated successes would be welcomed by policy makers and that the requisite program and policy modifications, which would become apparent as a result of these evaluations, would be readily accepted.

This rational implementation model proved to be grossly inadequate in its description of the realities that faced evaluators and decision makers. Problems were not easily defined, and proposed solutions were often vague or contradictory. Administrators believed that continued program funding depended on positive evaluation, and were accordingly suspicious, if not hostile. Evaluators failed to recognize that "evaluation is in many ways just another political act that occurs in a context where power, ideology, and interests are more powerful determinants of decision making than feedback about programs." (Cook, 1986, p. 200). This led to conflict between program managers and evaluators, weakening the usefulness of much evaluation work.

As the 1970s began, it became widely recognized that the evaluation of educational programs, especially those federal programs directed at the educational needs and problems of disadvantaged and minority students, was in need of substantial revision both at the theoretical and technical levels. In 1971 a National Study Committee on Evaluation, set up by the professional
education fraternity Phi Delta Kappa, announced that educational evaluation "was seized with a great illness" (Madaus, 1983; p. 14) and went on to encourage the development of new methodologies of educational evaluation and new training programs for evaluators.

The call for new ideas was enthusiastically joined by many evaluators; the late 1960s and early 1970s have been described as a period that was "vibrant with descriptions, discussions, and debates concerning how evaluation should be conceived" (Madaus, 1983; p.14). The same author goes on to point out that at that time, "the field of evaluation was amorphous and fragmented -- many evaluations were carried out by untrained personnel; others by research methodologists who tried unsuccessfully to fit their methods to program evaluations. Evaluation studies were fraught with confusion, anxiety, and animosity. Evaluation as a field had little stature and no political clout." (ibid, p.15)

Forces for Change

The science and practice of educational evaluation had to change, and it did. Over the next decade a variety of new models and approaches were developed to meet the weaknesses that had emerged. In particular, the new models sought to deal with two closely related problems: weak utilization of results, and narrow -- and often inappropriate -- methodologies. Towards the end of the 1970s, changes in Federal legislation and financing, in conjunction with rising public expectations regarding the accountability of schools for student achievement, significantly altered the focus and resources of evaluation efforts.

Improved Utilization

It has now generally been acknowledged that the appropriate role of educational evaluation is that of a management tool for the improvement of educational programs, not as a branch of pure or academic research (Eichelberger, 1986). This relatively recent insight is the result of the work done on the questions of utilization.

The assumptions of the 1960s and early 1970s, that evaluations would either be of prime importance in determining future program funding or be readily adopted on their merits, were proven wrong. Indeed researchers examining utilization found little evidence that evaluations played any significant role in decision making, either on program management or on questions of continuation or termination (Leuton, 1981, 1983; Cook, 1986). The earlier optimistic expectations came to be characterized as naive. One authority said that "the failure to use evaluation findings has almost assumed the proportions of a national scandal" (Guba, 1981; p. ix).

Attention to the usefulness of evaluation results gained initial impetus from the work of Michael Scriven (1967), who highlighted the important distinction between summative and formative evaluation. Summative evaluation, as he defined it, is "conducted after completion and for the benefit of some external audience or decision-maker (e.g. funding agency, historian, or future
possible users)." Formative evaluation, on the other hand, "is conducted during the development or improvement of a program or product" (Scriven, 1977, pps. 150-63). It assumes that the program is in a state of flux and that it will continue to evolve. It has as a primary purpose the generation of information that will be useful for the modification and/or improvement of an existing and ongoing program, and thus has acquired a management focus.

Whereas evaluation was normally a summative activity prior to the mid 1960s, the utilization issue forced a reconsideration of this position. Growing recognition that evaluation after the fact has little or no capability to contribute to program improvement led to an increasing emphasis on formative evaluation (see Patton, 1978; Cronbach and Associates, 1980; Stufflebeam et. al.,1971; Palumbo, 1984; McClintock, 1984).

The distinction between formative and summative evaluation also served to highlight the importance of the timeliness of information for decision making. This, in turn, led to increased emphasis on an integrated relationship between evaluators and decision makers defined, at least in part, by the utility of that relationship to the needs of the program manager. Thus the "front" and "back" ends of the evaluation process assumed greater importance in order to insure a role for evaluation in decision making. Pre-evaluation activities, such as evaluability assessment and other activities designed to promote a shared understanding of the purposes of the evaluation, are described as "front end." Those organization structures and incentives designed to encourage usage are understood to be a part of the "back end."

The roles of both evaluators and decision-makers began to change to meet the needs of integrated evaluation. Instead of serving as autonomous critics of the education system, evaluators moved towards membership in the education management team. Their commitment to usable evaluations increased, as did their sensitivity to the political dimensions of evaluation. Decision-makers began to accept the responsibility for defining decisions and decision-making processes, as well as for increased personal commitment to the evaluation process and to utilization of results Burry (1985).

The relevance of evaluation to the perceived interests of the program participants is clearly one of the key indicators of evaluation utilization, and the research indicated that the involvement of an administrator in the evaluation process is often the surest way to insure that relevance. Patton (1978; p.64) identified the "personal factor" -- that is "the presence of an identifiable individual or group of people who personally cared about the evaluation and the information it generated" -- as being especially important to the question of whether or not evaluations were used.

It was increasingly recognized as well that decision makers, especially within educational systems, must operate within a political, social, and economic environment in which decisions evolve from the complex and ongoing interplay of bargaining, compromise and negotiation. An autonomous evaluation poorly linked to users is, within this context, in danger of becoming little more than ammunition in a battle between factions.
As O'Reilly noted, results of evaluation research are more likely to be used by decision makers if the information is: (a) supportive of the outcomes favored by the decision makers and (b) does not lead to conflict among the set of relevant actors. He observed that:

Evaluation research, regardless of the rigor with which the study was conducted, is not likely to be regarded by decision makers as objective, nonpartisan information... the utilization of such information in decision making will probably reflect not any objective measure of quality of the research, but a number of factors independent of the evaluation study such as the degree of consensus or conflict among those involved in the decision process, the relative power of the participants, pressures on the primary decision makers, availability of other information, etc. (1981, p. 58)

The attempt to enhance utility has become one of the major themes supporting the development of new evaluative techniques. The importance of utilization is reflected in the current understanding that evaluation is most likely to be used if it serves to encourage and support program development -- and is not perceived to be simply a way of determining whether or not a program works or whether funding should be extended. The role of the evaluator has clearly shifted from that of a researcher to that of the manager of a process. In this new role, the evaluator's job is to work with decision-makers to identify needed information, collect and organize data, and then work with decision makers to determine the ways that the information can be used most effectively.

The utilization issue has thus prompted a thorough rethinking of the legitimate purposes of evaluation. It has been suggested, however, that the utilization issue has become salient, at least in part, because of methodological concerns. In the words of one evaluator, the failure to use evaluations "simply illustrates the poverty of traditional evaluations which are likely to fail precisely because they produce information that, while perhaps statistically significant, does not generate truly worthwhile knowledge." (Guba, 1981; p. ix).

New Methodologies

The prevailing evaluation methodology of the 1960s was built from the deductive research model of the social sciences. This approach had origins in agricultural and psychological research, in which controlled experiments were used to measure the effects of a treatment on a class of subjects. The treatment itself was based on theoretical models, ideally expressed as hypotheses regarding the causes and effects of treatment components. Sophisticated statistical modeling and measurement techniques were employed. The model had served well in other research contexts, and found strong support in the academic community -- where many evaluators resided.
Despite its scientific elegance, the model came under heavy criticism in the early 1970s, particularly from a Rand Corporation analysis of a collection of federal policies and social programs. This study sought to find out why some federal social change projects were more successful than others. The authors concluded, with reference to evaluations, that most "... ask the wrong questions and use the wrong measures. [The] study suggests that the problem with contemporary evaluation is deeply rooted -- that it stems from a fundamental lack of fit between the evaluation paradigm and the way that local projects operate and change takes place." (McLaughlin, 1980: p. 42).

The principal criticism was that the model failed to take into account the factors of principal importance in determining project effectiveness at local levels. The major findings regarding successful projects were that:

- **Implementation** dominated the outcome of planned change efforts;
- **Local factors** shaped implementation more than the adoption of a particular technology, the availability of information, or the funding level;
- **Mutual adaptation** between the project and the participants characterized implementation, with both changing over time.

These in turn depended on important local factors. Local administrative support; staff commitment, training and capability; site selection; and the broad-based participation of key stakeholders in the project were all valid predictors of project success.

But these factors were rarely addressed in evaluations based on the prevailing paradigm. The reasons lay in the nature of the paradigm itself:

... most project evaluations use an "impact" or "production function" model that correlates initial project inputs, such as level of funding and technology, to project outputs such as student achievement gains. This evaluation approach assumes that project treatment has a primary and direct effect on project outcomes. Local institutional factors and other variables that are important in the implementation process typically are omitted. Evaluators have lightly referred to these local factors as components of the "black box." The contents of the black box, it turns out, matter more to project outcomes than do the other factors that evaluators attempt to calibrate and assess. Insofar as unidentified causal variables have major and direct effects on project outcomes, an evaluation that fails to measure these variables can produce a specious finding of "positive program effects" or of "no significant difference." (McLaughlin, 1980: p. 42)
The model was thus weak in explaining the process components underlying implementation (Benson, 1987, p.43). This weakness made utilization of evaluations conducted this way difficult. Findings could not be trusted -- success or failure could not be attributed to the design or content of the program itself, so decisions to accept or reject a program could not firmly be made. Perhaps most importantly, the absence of insight into program implementation left the evaluations of little use to local implementors seeking to improve and refine the program.

The Rand study identified five lessons from research on project implementation with implications for evaluation methodology. First, treatment (or program) effects are indirect. For example, the impact of a new curriculum is mediated by the teachers, students and managers of the school in which it is implemented. Second, implementation choices dominate program outcomes. Third, implementation is a multi-stage, developmental process, in which the nature of the program changes. Fourth, implementors pursue multiple and often conflicting goals. And fifth, decisions made closest to the delivery level for services are most influential in shaping the way in which the program is actually delivered.

These findings applied broadly to a wide range of social change programs, including education change efforts. Other analysts pointed toward particular problems in using the input/output paradigm to evaluate educational programs. Much decision-making on the educational process must contend with the lack of a firm foundation of educational theory. It is difficult for educators to isolate the particular educational intervention that produces a given learning outcome. This inherent quality of schools causes design difficulties for evaluations based on a theoretical mode of the education process (O'Shea, 1981). As one scholar noted:

In seeking to evaluate an educational program, we may be led astray if we try to follow too closely the model of the scientific research study. An educational program is not a stable, natural phenomenon with built-in operating characteristics that people may discover and put to use, but not alter in any fundamental way. An educational program is a human artifact, highly complex, infinitely variable, and subject to incessant change. It defies precise definition or accurate measurement. The outcome of a precisely controlled scientific study of an educational program is almost certain to be either inconclusive or misleading." (Ebel, 1980: p. 228).

In sum, the predominant "input/output" paradigm was weak on two counts, both related to the "black box" of program processes. First, it failed to take into account local contextual and implementation factors, and research showed that these dominated program outcomes. Second, efforts to build theoretical models of the education process to establish the a priori structure of the black box, needed to generate the hypotheses on which measurement rested, were weakened by the absence of good educational theory. As a result, evaluation findings were difficult to use, either for program
continuation or for program improvement. The paradigm was weak for summative purposes, and of less use for formative decisions.

The implications for evaluators and evaluation sponsors increasingly committed to utilization of evaluation results were significant. Projects and programs could not be validly judged in isolation from the context in which they operated; nor did the implementation factors important to success lend themselves easily to modeling and quantification. Indeed, program implementation came to be seen as "... a complex, multi-stage process in which the phenomenon typically of interest to an evaluator -- the relationships between program inputs and system outputs -- is indirect and achieved against a noisy and constantly changing institutional setting." (McLaughlin, 1980: p. 13). Implementation is necessarily a heuristic process, one of learning and adjusting as opposed to a process of installation. Evaluation which fails to address these situational components will not develop information that has the immediacy and relevancy likely to be required by decision-makers.

Importantly, the earlier assumption that evaluations would lead to the identification of generalizable "truths" regarding the effects of educational interventions appeared untenable. The imprecise, evolving and site-specific nature of the implementation process meant that generalizations across many sites were risky. Moreover, even for a given site, evaluation findings, while useful at a given time, would decay as the program continued to evolve.

The direction in which evaluation models had to go was increasingly clear. Evaluation would become a tool for educational management, for both ongoing programs and the project-based interventions designed to improve quality. Formative evaluation designed to meet the information needs of decision-makers would be more useful than large-scale (particularly national) summative research-evaluation studies. These evaluations should take explicit account of context and process variables: this meant increasing attention to smaller evaluation studies at local levels. In the United States, this approach had the further potential advantage of generating information more quickly for use by the local decision-makers who indeed had influence over education programs. Such studies would be more likely to affect the choices of these decision-makers than large scale national studies which they believed, often correctly, did not necessarily apply in their local context.

Clearly, new models and methods of evaluation were needed to meet the new vision of the purposes of evaluation. In response to this need a radically different approach to evaluation appeared. Known by various names -- "qualitative," "naturalistic" -- this model explicitly addressed the need to measure process and context variables. And while it has not been adopted as a replacement for the input-output paradigm, it has influenced the models and current practice of educational evaluation in important ways, combining with more traditional measurement approaches to lead toward current eclectic evaluation practice.

Qualitative Approaches: Qualitative evaluation can be most accurately described as endorsing a particular point of view, as opposed to
encouraging any particular set of operating procedures and techniques. Indeed, given the very different philosophical base of qualitative methodology, it is important to bear in mind that the approach embraces a wide variety of measurement techniques. The point of view is derived in large part from the disciplines of ethnography and anthropology, and is distinctly different from the positivistic outlook on which the input/output paradigm is based. As a leading advocate the "naturalistic" approach has declared, "... its basic tenets are virtually the reverse of those that characterize positivism." (Lincoln, 1986: p. 29).

This approach is derived from a philosophical outlook that:

... emphasizes that groups of people can and often do have systematically different views of reality and consistently different ways of interpreting what actions, words, objects, and so forth mean. [This outlook also assumes] that people take action on their ideas and beliefs about the way the world really is, in light of their notions of what things mean, regardless of what is verifiable "scientifically" as "fact." (Dorr-Bremme, 1985: p. 66).

With this point of view, qualitative evaluators insist that the goal of evaluation must necessarily be more than just examining a program on-site. They maintain that in order to describe and understand, as completely as possible, the contextual framework within which any program operates, it is necessary to "get inside" and explain the program and its consequences from the point of view of the participants. Unlike the positivist evaluator, who seeks to control processes and outputs by the imposition of hypotheses, experimental designs and pre-determined measures of outputs, the qualitative evaluator seeks to study and understand a program and its context as it occurs naturally, without prior controls.

These techniques rely on observation and interviews, buttressed by document analysis. Unlike much conventional positivistic research which relies on data collection and measurement and pre-specified times (i.e., before-and-after treatment), qualitative approaches emphasize continuous data gathering, with periodic validity checks with the participants in the process being studied. Thus, for example, qualitative research on the implementation of an education reform would require that the evaluator participate in and observe the process, conduct structured or semi-structured interviews with key participants, and periodically prepare a summary analysis which identifies patterns and processes in operation. These summaries would be cross-checked with participants for validity. The evaluator would seek to reconcile any lack of consensus on the "truth" of the analysis through further data collection and a search for deeper explanations (see Middleton, 1985, for more detail on this approach).

Qualitative research approaches have gained in stature because of their ability to account for context and process in educational evaluations. A body of methodology has emerged to guide practice in education (Miles and Huberman, 1984; Guba and Lincoln, 1981), as well as in international development projects in a number of sectors (Salmen, 1987).
Used by itself, the qualitative approach has significant limitations. Most importantly, generalization beyond the particular process and context studied is difficult. Second, it requires well-trained and skillful evaluators, able both to participate closely in a process and to maintain the degree of objectivity necessary for analysis and interpretation. Finally, as a method it is open to partial and careless use (although positivistic methodologies are prey to the same problem).

Because of these problems, and because of its recent emergence, qualitative evaluation has not become widespread practice in U.S. educational evaluation. Its influence has been felt, however, as evaluators have moved towards hybrid models which seek to incorporate the best of both positivistic and qualitative methods in an attempt to marry valid measurement of outcomes, and assessment of process and context factors, to the need for localized evaluations which meet the utilization needs of educational decision-makers.

**Changes in Legislation and Funding: The Rise of Local Accountability**

As evaluators struggled with issues of utilization and methodology, changes also took place in the political and financing context of evaluation. Recent years have seen a lessening of support for the idea that the federal government is the appropriate agency for the solution of social problems, a perception brought about, at least in part, by the widely held view that many large-scale social programs substantially failed. This conservative national agenda has become evident as the federal government has increasingly shifted responsibility for both funding, and educational direction and evaluation management to state (SEA) and local (LEA) education agencies. As one evaluator noted "the golden age of evaluation, from a funding standpoint, is over" (Webster, 1987; p. 49). The evaluation capability that was developed on both the national and local levels during this "golden age" has, however, continued to nurture the continued growth within the profession. The new and smaller role of the federal government in education, and in evaluation, helps to explain the changes that have occurred within the past few years.

A central event in the transition towards increased local control of monitoring and evaluation was marked by the failure of the Vocational Education Data System (VEDS) (Hoachlander, et. al., 1985). Created by federal legislation in 1976, the system was to be built on State Occupational Information Coordinating Committees to improve the quality of labor market information; VEDS would provide a complex array of data on vocational education programs. The system imposed uniform requirements on the states for defining program types, student characteristics and financial information.

The system met with massive resistance and a relatively high level of non-compliance. The data requirements conflicted with established state and local practice, and placed an insupportable burden on state and local agencies. In the end, the system was withdrawn, and replaced in subsequent legislation with simpler requirements that reduced the data collection burden. At the same time, state-level systems for monitoring of vocational education remained comparatively undeveloped.
National data collection systems have been simplified. And evaluators are no longer faced with the problems associated with massive federal education programs and have turned their attention to activities at the individual school level. This reflects not only the accountability pressures and demands from both state and federal levels, and from communities, that are being placed on local school districts and schools, but also the lessons derived from extensive input-output research on education undertaken within the past several years. School effectiveness studies have clearly pointed to the importance of the educational experience at the local school and to the critical role played by the human resources there—the students and teachers (Murnane, 1982). This new interest in the educational program at the relatively smaller educational unit of the local school district or individual school, has increased the importance of monitoring and routine information collection at that level:

The object of traditional evaluation designs is changing, from a "programmatic intervention" focus to a focus on "schooling" itself. In other words, the "program" for evaluation focus is now the ongoing constellation of daily activities and outcomes constituting the program of the local school. This suggests the need for ... a considerable variety of information ... designed to facilitate any number of evaluative purposes from appraising the impact of specific programmatic interventions, to informing organizational and instructional planning and development activities, to, perhaps most importantly, monitoring the periodic health of the school work and learning environment. (Sirotnik, 1987, p.77)

The establishment of information bases at the school and district level has been supported by the rapid diffusion of computer technology and software.

The purposes of the collection and development of information have also changed significantly. Accountability has become the paramount concern at the local district level and the school principals are coming under increased pressure to produce demonstrable successes (see Fetler, 1986). As one evaluator described the current situation:

... the data for program improvement emphasis of the late sixties and seventies has been largely replaced by the accountability movement of the eighties. In the late sixties and seventies most school districts were involved in some type of curriculum development and had staff to do program improvement work; today, the level of curriculum development in school districts has dropped dramatically. Principals, as managers of school units are being held accountable for student achievement and consequently are being given more freedom in terms of the programs and strategies that they implement in their buildings. When
schools fail to produce desired results, principals are often moved instead of programs being abandoned. The evaluation or research emphasis is toward documenting the variables involved in "successful or effective schools" rather than toward documenting and improving successful programs. (Webster, 1987; p. 49).

This has resulted in a tremendous growth in the importance of testing. At the State level, these changes were increasingly reflected in the titles and responsibilities of evaluation offices. In the State of Washington, the Program Evaluation section became the section for Testing, Evaluation and Accountability; in Oregon, the evaluation unit is known as the Educational Program Audit Division. In at least two states (Oregon and Michigan), educational planning was subsumed within evaluation and testing units (Caully, 1981).

A major function of the evaluation units at the state level came to be the establishment and assessment of minimum standards for schools, with state funding tied to compliance, and for students, often with secondary diplomas contingent on passing a minimum competencies test. While not universal, these functions have become increasingly widespread as education reforms proceed.

Most evaluation units at the school district level are devoting a considerable portion of their resources to testing. For example, Webster (ibid; p. 50) reports that large school district evaluation units increased the proportion of their evaluation budget allocated for testing from 11% to 23.5% between 1978 and 1983. (Comparable figures for medium and small units were from 23.4% to 33.3% and from 19.0% to 30.3%.)

The resources directed to testing are necessarily being taken from other possible functions, and funds available for other aspects of evaluation have been reduced. For example, in Dallas, available federal evaluation resources dropped by 65% between 1978 and 1983 (Webster, 1987, p.37). The impact of these reductions has fallen mainly on local evaluations of process and curricula.

At the same time, data on student achievement at the school level are enabling researchers to choose successful schools for deeper inquiry into the factors of school effectiveness (Cohn and Rossmiller, 1987).

The current focus of on minimum standards, achievement testing, and monitoring is not without controversy. Many educators and evaluators are concerned that intense public and media interest in comparisons of test scores across schools, districts and student ethnic and racial groups distorts the meaning of test data. It is argued that achievement tests provide narrow measures of the broad goals and achievements of schools, and that short-term political pressures on schools are accompanied by little in the way of positive resources. Reduced resources for process evaluation make it more difficult for education managers to develop information-based strategies to overcome school deficiencies. One evaluator notes that his state provides no funding for remedial programs for the schools that do not meet standards.
(Bracey, in Caully, 1981). And, inevitably, accountability pressures from the state have led to conflicts between the state agency and local school districts.

Despite these problems, the accountability dimension of evaluation is likely to continue to dominate U.S. practice. Evaluators are now seeking ways to adapt their methods and approaches to this powerful factor in educational policy and decision-making.

The major adaptation has been general acceptance of the place of evaluation in school management, of the formative purposes of monitoring and evaluation processes. Thus the role of the evaluator has changed from that of the independent critic or researcher to that of a participant in the process of program improvement (Eichelberger, 1986: pp. 30-33).

Adaptation has also involved significant movement away from the rigorous standards of social science research towards more pragmatic approaches which provide timely information for decision-making. These approaches recognize that the users of evaluation are "... less concerned than academics with reducing the final few grains of uncertainty about knowledge claims, ... are more willing to trust their own experience and tacit knowledge for ruling out validity threats ... [and] they also expect to act upon whatever knowledge base is available, however serious its deficiencies." (Cronbach, cited in Cook, 1986: p. 219).

Summary

Educational evaluation in the United States began -- and has ended - - in schools and communities. Along the way an important excursion was taken into the realm of large scale research evaluation studies at the national level. The weakness of such studies in providing information useful to decision-makers seeking to improve education at local levels in a decentralized and diverse system of education led to their abandonment. In their place we now find eclectic approaches that meet the demands of utilization and accountability, include assessment of implementation processes, and emphasize achievement testing and monitoring systems that provide a continuous base of information on school operations. We also find parallel emergence of research centered on the factors that make schools -- not schooling -- more effective.

School accreditation systems, both voluntary and public, have provided an important level of institutional assessment. This, in turn, has helped protect the public against poor educational institutions and has helped administrators improve the quality of their schools.

The federal government played a crucial role in this process. Evaluation requirements and funding, tied to programs that financed educational change and improvement, made evaluation an increasingly integral part of educational practice. Federal funding initiated and sustained the development of evaluation capacity, through the establishment of State education agencies and the nurturing of the evaluation profession.
As evaluation matured, public reporting of achievement test results for schools and districts fed rising public concern about the quality of education. While the rise in accountability poses both technical and political problems for educationists and evaluators, it has undoubtedly energized the search for more effective schools.
LESSONS FOR THE IMPROVEMENT OF MONITORING AND EVALUATION IN DEVELOPING COUNTRIES

Assistance agencies and governments have a common interest in improved educational monitoring and evaluation. Better information on the outcomes and processes of education and training systems will contribute substantially to improved use of scarce resources, for the systems as a whole and for specific project-based interventions. It is crucially important to the incremental accumulation of knowledge that leads to better policy and practice for both; this is increasingly true as assistance strategies become more complex.

Effective investment strategies have yet to be established in the World Bank, either for project monitoring and evaluation or for the development of sustained government capacity. The failure to establish sustained capacity further weakens assessment of specific investments. In the World Bank, the combination of new lending strategies and constrained Bank resources is leading to more reliance on Borrowers for monitoring and evaluation: yet, in most cases, the capacity to do this effectively has not been developed.

Improvement in project monitoring and evaluation, as well as improved government capacity to manage the education system effectively, is dependent on the establishment of this capacity. Providing effective assistance to the effort is a significant challenge for the next generation of education investments.

Project monitoring and evaluation will continue to be important. The current trends toward improved and more widespread use of monitoring indicators at appraisal should be intensified, as should increased attention to administrative arrangements for M&E.

However, given limited resources, priority should be given to the establishment of sustainable capacity for system-wide monitoring and evaluation. The reasons are straightforward. First, without the support of institutionalized monitoring and evaluation capacity, M&E for specific projects will continue to be weak, regardless of design improvements. And second, sector and policy-based lending can only be monitored and evaluated effectively through strong and sustained government action.

Strategies for Building M&E Capacity

To build sustained monitoring and evaluation capacity, strategies will be needed to address five central questions. Who will be served by M&E? For what purposes? What information is most needed? What kinds of M&E systems are needed to capture this information? How can such systems be established and sustained?

The U.S. experience suggests some initial answers. Some of these will be familiar to development educators; others may be less so. In all
cases, these should be treated not as prescriptions, but as guidelines to be explored and tested through lending operations.

**Who Should be Served?**

*Policy and decision-makers at all levels whose decisions affect delivery of services, and thus quality and efficiency: this includes teachers, headmasters and district staff.*

The U.S. experience is unequivocal on this point. Monitoring and evaluation should serve the needs of policy and decision makers in a position to improve educational practice. Of considerably less importance are policy and decision makers with only peripheral ability to influence the education enterprise and researchers interested in generating universal truths about the education process.

In the decentralized and diverse U.S. system of education and training, these decision-makers are found in communities, schools, districts and state education agencies. The predominance of local and state financing for schools places effective policy and implementation decisions at these levels. The role of the Federal Government is restricted to the provision of fiscal incentives intended, often with success, to pull the practice of education in a given direction over a longer term.

In a number of larger developing countries (India, Brazil, Nigeria) education is organized similarly, with roughly analogous decision-making structures (although invariably with a stronger Federal role). In smaller countries, in many of which the entire school systems are smaller than U.S. school districts, equivalent decentralization has not emerged and may not be necessary.

However, in all cases, U.S. experience demonstrates the crucial role of implementation in determining the quality and effectiveness of both ongoing programs and interventions aimed at improved practice. Schools, and classrooms, are the places where education is delivered and improved, where the day-to-day decisions which most immediately affect quality and efficiency are made. Thus headmasters and teachers, and communities, need not only to provide, but also to use information about the processes and outcomes of education.

The centralized monitoring and evaluation systems that the World Bank has (unsuccessfully) sought to implant have ignored decision-makers at lowest levels. While the Bank and governments will continue to need aggregate data at national levels for project and system M&E, systems should be designed to provide information to schools and districts on a routine and short-term basis. This will require "front-end" design attention to the scope and information needs of decision-making at different levels of the system. Implementation will require training and technical assistance to lower-level decision makers not only in data collection, but also in data use. This will be, in most cases, a new kind of activity for the Bank and for Borrowers.
For What Purposes?

To generate information useful in improving educational practice at the school level.

The relatively small impact on the schools of national-level summative research evaluation studies in the U.S. has been well recognized. Current practice (and theory) emphasizes the formative use of monitoring and evaluation information by decision-makers in a position to act incrementally to improve practice.

Large-scale summative evaluation has, so far, failed on two dimensions. One is utilization. Recognition that implementation conditions vary from site to site, and of the long gestation period for large studies, has tended to make findings irrelevant at the level where decisions are taken. Decision-makers at local levels have been reluctant to accept findings that are based on national samples, whether relevant or not, unless they support directions for policy and practice determined primarily by local factors. A second is theory failure: education has yet to evolve a generally accepted theory which enables decision-makers in diverse contexts to interpret and apply general findings with confidence.

This is not to argue that educational research should cease. Indeed, it is of considerable importance that it be strengthened. The move towards research at the school level in the U.S. indicates the continuing vitality of such inquiry.

However, it does suggest that priority should be given to the purpose with the most likely immediate payoff -- improvement at the school level.

Moreover, a clear and strong assistance agency policy position in support of formative evaluation might help reduce the resistance of governments to "evaluation" that is seen as judgmental of their efforts. Such a policy position would be reinforced by designs which explicitly incorporate analysis of the decision-making needs at different levels.

Donors and governments will continue to make summative judgments regarding the effectiveness of investments. These could be significantly strengthened over current practice by the information generated for formative purposes. Summative evaluations of educational outcomes have, in any case, played almost no role in project evaluation. Any special studies that are contemplated should be cast in a formative mold, and they ought to be centered on school effectiveness.
What Information is Needed?

Educational and economic outcomes, costs, and the implementation performance of schools

In current U.S. practice, educational achievement data hold pride of place: equity concerns lead to collection and reporting of the data not only by school and district, but also by socio-economic characteristics of students. These data are supported by a range of management information on enrollments, staffing, costs, facilities, and so on. Implementation processes are assessed through surveys, observation, qualitative analysis, tacit knowledge, and, less frequently, special studies. Accreditation studies provide a basis of formative information on overall institutional quality.

Information on the economic outcomes of general education is not routinely assembled at the institutional level; perhaps because of the failure of the VEDS system, there is also very little information on the employment outcomes of vocational training.

The absence of data on educational and economic achievement is the most glaring weakness in most developing country contexts, not only for project institutions, but for the system as a whole. The Kenya example cited earlier provides one of the few exceptions. Data on the unit costs of education of various kinds is equally weak. Management information systems have a spotty history, and information on implementation at the school level is exceedingly rare.

The consequences for governments and the World Bank have been discussed at length. A great deal of information is available on inputs and aggregated project costs; very little is available on the factors that permit learning about the effectiveness of schools, projects or systems.

Much of this is attributable, of course, to the general failure to address crucial data needs in the design of M&E investments, and to the failure of most investments in M&E capacity. In particular, the failure to establish achievement testing systems in many countries has made it impossible to assess the educational effectiveness of investments. The attempt to establish tracer studies of employment outcomes has similarly failed in most countries. Declining levels of Bank supervision, larger and more complex projects and weak inspectorate systems have combined to reduce knowledge of implementation processes.

More success in developing the right kinds of information will depend on much stronger designs for M&E systems that address the need to specify decision-makers and their needs at different levels, emphasize the formative purpose of monitoring and evaluation, and give priority to data on achievements, costs and implementation processes.
What Kinds of Systems Are Needed?

Testing, management information and inspection that provides information useful at all levels of the system.

U.S. systems have become eclectic. The foundation is achievement testing designed and administered by states, districts, schools and teachers. The complex and variable system of testing is dictated by the decentralized control of curricula and school financing. Testing systems have in common, however, a comparatively high degree of professionalism, nurtured in large part by the well-developed educational research community in universities and test development and publishing agencies.

The first, and cornerstone element of an effective M&E system is achievement testing. The development of testing systems is complex, involving an array of technical and organizational issues (see Heyneman, 1987, for a good summary of the issues for developing countries). It is sufficient for this paper simply to note that systems which assess student learning, and which are capable of reporting the data at the level of schools and the next higher administrative level (as well as at national levels), and which permit identification of trends in achievement over time are central to effective monitoring and evaluation. Without them, other kinds of information are of little use in improving the quality of education, or in making choices between alternative strategies and investments.

Management information systems which capture the allocation and flow of resources and students at school, district and national level are equally important. Without knowledge of costs, teacher qualifications and deployment, curricula, and student qualifications and flows, it is impossible to begin the task of identifying patterns which are associated with higher achievement. Much of this information is also essential to system administration.

To be useful, this information must, obviously, be assembled and reported routinely and in a timely manner. The key is to keep these requirements simple. This takes not only micro-computers (which can be very helpful) but also attention to the skills and motivation of the individuals whose action (or lack of it) determine whether the data is accurate and timely. Monitoring and reporting must be integral parts of the job descriptions of key actors (headmasters, district officials), and training provided. Performance in management information tasks should be important in employee evaluations. Outstanding performance should be rewarded; poor performance a cause for disciplinary action.

Qualitative assessment can play two important roles. One is in the identification of educational issues and problems as seen by communities, students and school staff. A second is in assessment of the operation of schools, both for routine service delivery and implementation of change. In both roles it is an important companion to statistical reporting. Qualitative assessment can be handled through contracts with strengthened university research units or the inspectorate (with special training) on a routine basis. These studies are most likely to be practical when conducted on a sampling
basis. Qualitative assessment would be a good candidate for the "special studies" that assistance agencies routinely finance.

Development of the crucial human aspect of management information systems is much more difficult in developing country contexts than provision of equipment and the design of forms. There are many barriers to success: overworked and underpaid staff, competing priorities, lack of incentives, fear of delivering bad news. Unless these factors can be squarely addressed in the context of a formatively oriented system which provides information that educators can use, investments in management information -- and indeed in M&E as a whole -- are unlikely to succeed.

The third essential component of an M&E system is school inspection. In many countries, the inspection function (assessing compliance with minimum standards) is integrated with the provision of professional support to school staff. These roles can conflict, with inspectors seeking at once to be evaluatory and supportive. Standards are often lacking, and meaningless in the context of constrained resources.

It may be appropriate in some countries to separate professional support from the inspection function, with the latter moving toward a periodic accreditation assessment. The regional and voluntary nature of U.S. accreditation is likely to be relevant only in larger countries, but a separate accreditation function could make important contributions (as has been the case in Korea).

The importance of these three system components is well understood, even if little success has been encountered in putting them in place. An improved record of accomplishment ultimately rests less on the details of system design (though these are important) than on success in establishing any system as a sustained and integral part of the educational enterprise.

How Can Monitoring and Evaluation Systems Be Established and Sustained?

Through comprehensive institutional development strategies in sector operations targeted on strong existing administrative structures and on primary education, with a strong policy and incentive framework, central units with strong components at lower levels, institutional structures that highlight M&E, strong investment in staff socialization and development, and a long-term incremental approach.

This is principally an institutional development question. The fact that monitoring and evaluation systems are the object of the effort is relevant only to the extent that these are considered peripheral -- or threatening -- to more important institutional concerns. Hence the experience of the U.S., and its potential application in developing country contexts can profitably be examined with institutional development concepts. The answers to this question will thus be developed by outlining a conceptual framework for institutional development, using it to interpret U.S. experience, and
finally, identifying implications for investment strategies intended to develop institutionalized monitoring and evaluation capacity.

A Conceptual Framework for Institutional Development

Arturo Israel's recent analysis of institutional development in developing countries provides a useful framework for such an analysis (1987). The main components of the framework can be summarized as follows.

Institutional development depends fundamentally on incentives for performance of the individuals in the institution. The degree to which incentives are inherently present in turn depends on the nature of the activities carried out by the organization. Some activities are intrinsically stronger with respect to performance incentives than others, and consequently, more likely to lead to institutional development.

The strength of inherent incentives depends on two factors in the work of the institution. The first is the degree of specificity of the activities in which the institution is engaged. Specificity is high when objectives, means and ways of controlling performance can be clearly identified; and when the effects of these activities are intense, immediate, identifiable and focused. Under such circumstances the performance of individuals, and hence the potential utility of incentives in motivating that performance, is relatively high. In contrast, specificity is low when objectives, means and controls can be less clearly determined, and the effects are weaker, delayed, less identifiable and diffuse. Under such conditions intrinsic incentives are much weaker.

The second is competition. Defined more broadly than economic competition, the concept includes "surrogates" for market competition: external pressures from clients, beneficiaries and suppliers; external pressures from political or regulatory bodies; and internal competition among individuals or units within the institution. Again, intrinsic incentives are higher under conditions of high competition.

Israel's review of World Bank experience indicates that activities in high technology, finance and industry are generally high in both specificity and competition. Activities with low technology, and those relating primarily to human behavior ("people-oriented" activities), are low in both. Institutional development has been more successful where specificity and competition are high. In his analysis, education is an exemplar of a low technology, people-oriented activity where institutional development has been relatively unsuccessful.

Israel suggests that priority for institutional development work should be given to low specificity, low competition institutions. Strategies for institutional development under such conditions operate at both national and institutional levels.

At the national level, increasing awareness of the need for improved institutional performance, simplifying objectives and programs to match institutional capacity, reducing the distortions in wages and salaries that
discourage improved performance, and building national commitment to institutional development are the crucial factors.

There are two strategy components at the institutional level. One is made up of general strategies: a) simplification of objectives and activities at the institutional level; and b) increased professionalization of "people-oriented" work through socialization. At this level, for example, monitoring and evaluation should be restricted to the minimum amount of quantitative and written information; and effective oral communication between the levels of the system improved.

The second element consists of additional management action to simulate specificity and competition. For specificity, these include inter alia improved employee selection and recruitment; staff incentives that include both pecuniary and normative components; staff evaluation systems that permit tracing of individual staff actions; stronger middle management support; more complex organizational structures with more levels and stronger "staff" structures. Means for increasing competition include giving greater voice to clients, beneficiaries and suppliers; improved performance in political and regulatory agencies; and enhancing internal competition by focusing management on results. Internal monitoring and evaluation plays a crucial role in internal competition, "... but its use should be selective, more economical, and driven by the interest of managers." (1987, p. 175).

The U.S. Experience Interpreted

Incentives have played an important part in establishing M&E capacity in the U.S. At a national level, by tying funding and requirements for evaluation to prestigious programs of educational improvement, the federal government created not only a broad context that associated M&E with educational change, but also provided the means through which institutional capacity could be established. This is seen most directly in funding for the creation of evaluation capacity in state education agencies.

Federal funding also supported the "professionalization" of the practice of evaluation. From this profession came the methodological modifications that enabled evaluation practice to move more closely to the needs of decision-makers, and away from large scale national research.

The administrative structure for M&E that has emerged is complex, with federal standards and guidelines, coordination and implementation from state agencies, and implementation at local levels. Staff support for the function has expanded at all levels. Vertical coordination has been established by mandated reporting requirements.

These requirements, as well as the minimum standards for student achievement established at the state level and criteria for accreditation have increased the specificity of the work of educational evaluators, district superintendents, headmasters and teachers.

At the same time, it is important to note the failure of the Vocational Education Data System in the first half of the 1980s. An overly
complex monitoring and evaluation system was imposed at a national level before basic monitoring capacity existed at the state and local levels.

Competition among units has increased as accountability pressures from the public led to reporting of achievement test scores at school and district levels. While not universally favored by educationists, these pressures and data have strengthened attention to improvement at the school level. Indeed, with reduced federal funding for evaluation, these pressures may be crucial to maintaining the current level of M&E capacity.

These developments took place against a backdrop of already strong educational institutions at all levels, at least in comparison with developing countries. Awareness of the need to strengthen the capacity of schools was high, and grew higher as public pressure for accountability increased. While salaries in the education sector were thought low in comparison with other professions, there were clear rewards for longevity and good prospects for advancement for competent individuals.

Hence in broad terms, the institutional development history of the U.S. conforms to a number of the elements of Israel's framework. This provides support for the idea that U.S. experience can provide indications of new directions for investment in monitoring and evaluation.

Implications for the Design of M&E Systems

Investment in monitoring and evaluation capacity should be targeted. Not every developing country will have the multi-level administrative capacity that is important to effective monitoring and evaluation, nor will every country be willing to make the strong and sustained policy commitment that system development requires (in the U.S. this took more than a decade). Given limited resources, an appropriate general strategy would be to target a relatively small set of countries, perhaps one per region, where administrative capacity and policy commitment is relatively strong. Assistance operations in these selected countries could serve to test strong and well-financed investment designs.

Priority should be given to primary education. Given the importance of improving primary education, and the potentially strong role of monitoring and evaluation in improving school quality, this sub-sector might benefit most substantially. Beginning with primary schools has the advantage of comparatively simple and standardized operations. Once established, the system could be expanded to more complex sub-sectors, such as training, in which monitoring of employment outcomes is essential and which is often dispersed across several ministries and agencies. At the same time, an opportunity to develop M&E for a training system with strong management and policy commitment should not be ignored.

Priority should be given to sectoral operations. These tend to be undertaken where management and policy commitment has been relatively well established. The need for government capacity is arguably greatest where sector lending has become the rule. Finally, the policy-based nature of these operations provides an appropriate context for establishing the policy and
incentive framework central to developing institutional M&E capacity. In particular, the opportunity exists to tie requirements and funding for monitoring and evaluation to support for sub-projects, including assistance to intermediate-level administrative units. This approach, however, will require a relatively high level of preparation by both governments and donors for the M&E component.

Establish a policy and incentive base. Perhaps the clearest policy lesson from the U.S. is the effectiveness of attaching requirements, standards and funding for monitoring and evaluation to program support for educational change and improvement. It is important to note that this is not the same as attaching such requirements to routine funding. By associating M&E with broader programs of educational change, it has been possible to increase both the level of specificity of activities and introduce elements of internal competition within systems.

Legislation may be required to establish testing requirements and capacity, especially when separate (and relatively autonomous) testing agencies are to be established. Legislation, or changes in administrative rules, may be required to facilitate contracts for services between M&E units and university-based researchers and evaluation consultants.

Clear policy statements from the Ministry on the formative purposes of M&E activities, supported by training and orientation at all levels to give the statements life, is important to overcoming the inherent resistance to "evaluation" that has defeated previous efforts.

Legislation (such as to establish minimum student standards) may not be necessary, except in a few upper middle income countries with well developed systems generally, and possibly where control and financing of education has been substantially decentralized to states or provinces.

Build a central capacity with strong vertical links to intermediate administrative units and schools. The "center" should be defined as that point in the system where effective control over curriculum and financing rests. In small countries, this will be the ministry of education; in some larger countries, these responsibilities may have been decentralized. In this latter case, strategies may well benefit by phased implementation, with one or two states providing an initial test ground for systems development.

Building capacity at intermediate and school levels is essential. Both the failure of previous World Bank strategies, and the success in the U.S., argue strongly for a comprehensive program that incorporates all levels from the beginning. This will require a much higher level of "front-end" analysis of decision-making patterns and information needs than has been thus far attempted. The higher level of preparation work required is another argument for selective targeting of countries.

Institutional structures should highlight the M&E function. World Bank experience with attempts to build M&E capacity as a subordinate part of educational planning, or within Project Implementation Units (PIUs) or construction agencies, has not been encouraging. Locating M&E in specialized
curriculum and research institutions, in contrast, has been successful in Korea. In the U.S., M&E units at the state level, including those that subsume planning have been effective. A partial explanation for the relative effectiveness of units designed specifically for M&E lies in the higher degree of accountability that can be required, as well as in the potential for greater autonomy of action.

Develop the human resources for M&E. There are two goals for human resource development in support of monitoring and evaluation systems. One is to address needs for professionalization of the field, and socialization of the system to support the monitoring and evaluation effort. The second is to develop the technical and managerial competencies of actors in the system at all levels.

The professionalization dimension has been largely overlooked in developing countries, yet it has been an important element in establishing the U.S. system and holds an important place in institutional development generally. A number of instruments are available. Experience suggests that competitive salaries and work conditions rank very high in attracting strong individuals to and establishing status for a profession. Government willingness to establish these conditions should be a principal test of the level of policy commitment, and thus of the feasibility of an M&E investment. Other instruments include support for journals and attendance at international conferences.

University-based education researchers are an invaluable resource to monitoring and evaluation systems. Investment in the social science and education faculties of universities should be seen as an important linked component for M&E investments, one which supports professionalism and the development of technical competencies. As noted, administrative barriers to effective use of this resource (through consulting and contracting) may have to be removed.

Developing technical competency among the staff of the education system is of course necessary; indeed, technical assistance, fellowships and in-country training have been the modal inputs to building M&E capacity in World Bank projects. To develop sustainable institutional capacity, however, these inputs should be used to support a staff development program at all levels of the system. This program, in turn, must be derived from the task requirements of staff at different levels and with different roles. Broad-based, general training not tied to specific roles and functions is much less likely to contribute directly to system development.

Management training is also important. Much of the challenge of monitoring and evaluation is management of both people and information. Without competent management, the most technically sophisticated evaluation staff will not be able to make M&E systems work.

In-service training should be continuous. Emphasis should be given to training across levels of the system to enable individuals whose interactions are essential to system success to understand their roles in relation to each other.
Finally, internal communication systems should be developed that permit information to flow both ways in the system. Participants in the system (headmasters, district officials) need to know that their inputs are being used. They need also to be able to bring issues and problems to the attention of higher level officials, and get response. Communication of this type (through newsletters, internal circulars, frequent seminars) contributes importantly to the socialization and professionalization of staff, and permits system improvement from the ground up.

**Strategies should begin with the basics, and evolve incrementally toward increased sophistication.** Lessons from experience indicate that fully developed monitoring and evaluation systems should serve decision-makers at all levels with information useful in improving educational practice in schools. Information on outcomes, costs and implementation at the school level is needed from testing, management information and inspection systems. Establishing these systems requires institutional development strategies with strong policy and incentive frameworks, strong central units linked vertically to schools, institutional structures that elevate the status of M&E, and broad and continuous programs of staff development.

This is a very large agenda. Except under the most favorable conditions, it will need to be addressed in an incremental fashion over the course of several investments.

The failure of the VEDS system in the U.S., and Israel's conclusions about monitoring and evaluation in general, indicate the need to keep data requirements at a minimum to reduce the burden of new tasks on actors in the system and, thus, to improve the chances of implementation.

This principle can be extended to cover the M&E system as a whole. In initial stages, all the basic system components should be put in place -- testing, management information, inspection -- and at all levels -- central, intermediate, school. But the level of technical sophistication of each component should be kept as low as possible. This strategy provides a core around which institutional development activities (policy, incentives, training) can be organized. It permits actors in the system to learn their roles and how to work together. And not least, it provides an immediately stronger capacity to support project monitoring.

When the basic system is established, it can be expanded incrementally along several dimensions. One is to extend coverage to more components of the education and training system. Another is to increase the technical sophistication of processes (such as testing, or management information systems). A third is to expand the range of decisions supported by the system, and hence the scope of data collected.

**CONCLUSIONS: OPERATIONAL ISSUES FOR THE WORLD BANK**

The generally unsuccessful experience of the World Bank and the substantial achievements in the U.S. demonstrates that the development of
sustained and effective educational monitoring and evaluation requires strong policy and funding commitment over a relatively long period. Systems must be designed to support decision-making at various levels, providing information useful in school improvement. Effectiveness depends crucially on professionalization of staff and on technical and managerial competence.

The preparation of such investments (as part of sector or project operations) requires a level of analysis and design work beyond that normally made available through World Bank or government resources. Moreover, the policy and technical knowledge required is in extremely short supply in both. This gap between need and resources can be partially addressed by targeting a few countries at a time. However, this will require additional staff expertise in the Bank. Additional specialized consultant resources will be required as well.

The investment costs of M&E system development are impossible to estimate precisely, but they will not be insignificant. If World Bank lending for human resources expands, these will be relatively easily absorbed. If it does not, investments in M&E capacity will represent trade-offs with other pressing investment priorities.

In either case, increased policy commitment in the Bank to building M&E capacity is a sine qua non to progress. In this regard, recent experience is not encouraging. The current Bank policy statement on project monitoring and evaluation (Operational Manual Statement 3.55) is more than a decade old, and needs updating. A revised and stronger policy statement on monitoring and evaluation generally was developed by an inter-sectoral committee just prior to the reorganization of the Bank, but was not approved during the transition period. This statement needs to be revived, although it is not yet clear where in the new Bank overall responsibility lies for such efforts.

Increased effort in education lending alone is of course possible, but this will be possible only to the extent that senior management provides clear policy support, incentives for staff effort and the additional resources required.
BIBLIOGRAPHY


Bogdan, Robert and Sarı K. Biklen, Qualitative Research for Education: An Introduction to Theory and Methods, Allyn and Bacon, Boston, 1982.

Bogdan, Robert and Sari K. Biklen, Qualitative Research for Education: An Introduction to Theory and Methods, Allyn and Bacon, Boston, 1982.


Minnesota Research and Evaluation Center, Minneapolis, Minnesota, June, 1984, pps. 27-40.


<table>
<thead>
<tr>
<th>Title</th>
<th>Author</th>
<th>Date</th>
<th>Contact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recent Developments in Commodity Modeling: A World Bank Focus</td>
<td>Walter C. Labys</td>
<td>October 1988</td>
<td>A. Daruwala 33716</td>
</tr>
<tr>
<td>Public Policy and Private Investment in Turkey</td>
<td>Ajay Chhibber, Sweder van Wijnbergen</td>
<td>October 1988</td>
<td>A. Bhalla 60359</td>
</tr>
<tr>
<td>Commercial Bank Provisioning Against Claims on Developing Countries</td>
<td>Graham Bird</td>
<td>October 1988</td>
<td>I. Holloman-Williams 33729</td>
</tr>
<tr>
<td>Import Demand in Developing Countries</td>
<td>Riccardo Faini, Lant Pritchett, Fernando Clavijo</td>
<td>November 1988</td>
<td>K. Cabana 61539</td>
</tr>
<tr>
<td>Export Supply, Capacity and Relative Prices</td>
<td>Riccardo Faini</td>
<td>November 1988</td>
<td>K. Cabana 61539</td>
</tr>
<tr>
<td>The Effects of Financial Liberalization on Thailand, Indonesia and the Philippines</td>
<td>Christophe Chamley, Qaizar Hussain</td>
<td>October 1988</td>
<td>A. Bhalla 60359</td>
</tr>
<tr>
<td>Linking Development, Trade, and Debt Strategies in Highly Indebted Countries</td>
<td>Ishac Diwan</td>
<td>November 1988</td>
<td>I. Diwan 33910</td>
</tr>
<tr>
<td>Public Finances in Adjustment Programs</td>
<td>Ajay Chhibber, J. Khalilzadeh-Shirazi</td>
<td>December 1988</td>
<td>A. Bhalla 60359</td>
</tr>
<tr>
<td>Women in Development: Defining the Issues</td>
<td>Paul Collier</td>
<td>December 1988</td>
<td>J. Klous 33745</td>
</tr>
<tr>
<td>Title</td>
<td>Author</td>
<td>Date</td>
<td>Contact</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>-------------------------------</td>
<td>---------------</td>
<td>---------</td>
</tr>
<tr>
<td>WPS131 Implementing Direct Consumption Taxes in Developing Countries</td>
<td>George R. Zodrow, Charles E. McLure, Jr.</td>
<td>December 1988</td>
<td>A. Bhalla 60359</td>
</tr>
<tr>
<td>WPS134 International Trade in Financial Services</td>
<td>Silvia B. Sagari</td>
<td>January 1989</td>
<td>W. Pitayatonakarn 60353</td>
</tr>
<tr>
<td>WPS135 PPR Working Papers Catalog of Numbers 1 to 105</td>
<td>PPR Dissem. Center</td>
<td>November 1988</td>
<td>Ann Van Aken 31022</td>
</tr>
<tr>
<td>WPS136 Pricing Commodity Bonds Using Binomial Option Pricing</td>
<td>Raghuram Rajan</td>
<td>December 1988</td>
<td>J. Raulin 33715</td>
</tr>
<tr>
<td>WPS137 Trends in Nontariff Barriers of Developed Countries: 1966 to 1986</td>
<td>Sam Laird, Alexander Yeats</td>
<td>December 1988</td>
<td>J. Epps 33710</td>
</tr>
<tr>
<td>WPS138 Fiscal Adjustment and Deficit Financing During the Debt Crisis</td>
<td>William R. Easterly</td>
<td>January 1989</td>
<td>R. Luz 61760</td>
</tr>
<tr>
<td>WPS139 A Conceptual Framework for Adjustment Policies</td>
<td>Bela Balassa</td>
<td>January 1989</td>
<td>N. Campbell 33769</td>
</tr>
<tr>
<td>WPS140 Building Educational Evaluation Capacity in Developing Countries</td>
<td>John Middleton, James Terry, Deborah Bloch</td>
<td>February 1989</td>
<td>C. Cristobal 33640</td>
</tr>
</tbody>
</table>