EVALUATION AND EDUCATIONAL REFORM

Policy Options

Education Policy in Latin America and the Caribbean

TECHNICAL REPORT 3

U.S. Agency for International Development • Bureau for Latin America and the Caribbean
Office of Regional Sustainable Development • Education and Human Resources Division

EVALUATION AND EDUCATIONAL REFORM
Policy Options

Benjamín Álvarez H.
Mónica Ruiz-Casares

EDITORS

1998

Education Policy in Latin America and the Caribbean

The revitalization of education, a common goal for all the nations of the Americas, cannot be achieved without the participation of all their citizens and the contribution of available technical knowledge. The Technical Paper series aims at opening new venues to facilitate broad regional dialogue on educational policies, and develop useful learning on their implementation and results. It also hopes to provide opportunities for sharing comparative ansalyswes, evaluations, and research perspective for improving public and private educational decisions. The Latin America and Caribbean Bureau of the U.S. Agency for International Development expects to contribute with the Technical Papers to the efforts made by several national and international organizations committed to the improvement of education throughout the hemisphere.

The ABEL Project

The Advancing Basic Education and Literacy Project is operated by the Academy for Educational Development with subcontractors Creative Associates International, Inc., Education Development Center, Florida State University, Harvard Institute for International Development, and Research Triangle Institute. The ABEL Project is funded by the Center for Human Capacity Development, Global Bureau, U.S. Agency for International Development (USAID).

This document was produced by the Academy for Educational Development with funding and guidance from USAID's Bureau for Latin America and the Caribbean and the Global Bureau's Center for Human Capacity Development. The findings, conclusions, and recommendations expressed in ABEL documents are the authors' and do not reflect the opinions of any of the institutions associated with the ABEL Project or USAID.

Project No. 936-5832
Contract Nos. HNE-C-00-94-00075-00, HNE-Q-00-94-00076-03

CONTENTS

Foreword v
Abbreviations and Acronyms vii
Introduction ix

The Contribution of Evaluation to the Learning of Nations 1
Benjamín Álvarez H.

SECTION I: NATIONAL SYSTEMS

Monitoring National Educational Performance 23
Thomas Kellaghan

Evaluation and Curriculum Standards Indicator Systems in an Era of Educational Reform 61
Gilbert A. Valverde

International Monitoring of the Goals of Human Development: The Case of the Secrectariat Pro Tempore of the Americas 93
María Inés Cuadros

SECTION II: LESSONS OF HISTORY

Social Impact of Educational Performance Evaluation Systems: The Case of Chile 115
Erika Himmel

The System for Evaluating the Quality of Education in Colombia 145
Gabriel Restrepo

SECTION III: TEACHER EVALUATION AND PROFESSIONALISM

The Evaluation of Teachers 171
Carol Dwyer

Evaluating Teacher Performance in Latin America 203
Franciso Álvarez Martín
in collaboration with María José Álvarez and Paula Vergara

SECTION IV: EVALUATION OF THE ORGANIZATION OF EDUCATION

Evaluating the Performance of Individual Schools 237
William J. Webster and Robert L. Mendro

Monitoring and Evaluation of Educational Reform Initiatives in the State of Paraná, Brazil 293
María Teresa de la Fuente, Heloisa Luck, and Corinna Ramos

Building a State Evaluation System: The Experience of the State of Aguascalientes, Mexico 317
Margarita María Zorilla Fierro

CONCLUSIONS

Challenges and Policy Options for Educational Evaluation 333
Benjamín Álvarez H. and Ray Chesterfield

APPENDIX

International Perspectives on Standards and Assessment: A Selected Bibliography . 345
Teresa Kavanaugh

Contributors 353

FOREWORD

In recent years, standards and assessment in education have come to the forefront of national and international development dialogue. The catalyst fueling this dialogue is the compelling need governments feel to prepare their citizens for life in the information society of the twenty-first century, driven by a competitive global economy. As questions of global competitiveness are more closely examined, improving the quality of education systems and the human capital they produce is an imperative. And questions of how to improve the education system inevitably lead to the discussion of standards and assessment.

While a majority of people may agree that improvement in instructional quality and student learning outcomes is necessary in most nations of the hemisphere, it is not so easy to gain consensus on how the information will be used or determine whether it adequately measures educational efficiency or performance outcomes. The current debate in the United States over the adoption of national standards is a prime example of the difficulties surrounding this subject.

In recent decades, there have been many projects aimed at improving education statistics in Latin American countries. These efforts have greatly improved availability of technology but have not provided relevant guidance on the development of appropriate statistics, indicators, and information systems. The USAID/LAC Bureau sincerely hopes that the studies in this volume contribute significantly to the information base available to practitioners charged with developing standards and assessment mechanisms and policy makers responsible for making decisions about the use, dissemination, and implementation of education standards and assessment instruments. We also hope this report will begin to fill the information gap as well as lead to further research on the topic.

On behalf of the U.S. Agency for International Development and the Bureau for Latin America and the Caribbean, we dedicate this volume to school children throughout the hemisphere and offer our sincerest thanks to the authors for their outstanding level of effort on this project. Particular praise is due to the Academy for Educational Development for the excellent work and dedication of such individuals as Francy Hays, Mónica Ruiz-Casares, and Dr. Benjamín Álvarez for their intellect and leadership in examining the complex issues of education policy reform and standards and assessment.

Sarah Wright

Education and Human Resources Team
Bureau for Latin America and the Caribbean
U.S. Agency for International Development

LIST OF ABBREVIATIONS AND ACRONYMS

ABEL	Advancing Basic Education and Literacy project
ACP	Assessment of Course Performance (*United States*)
AERA	American Educational Research Association
APA	American Psychological Association
CEMIE	Multinational Center for Educational Research (*Costa Rica*)
CEPES	Paraiba Centers for Solidarity Education (*Brazil*)
CIDE	Center for Educational Research and Development (*Chile*)
CIPP	Context, Input, Process, Product model
CONSED	National Council of Secretaries of Education (*Brazil*)
DIP	District Improvement Plan (*United States*)
ETS	Educational Testing Service (*United States*)
FADE	School Development Support Fund (*Brazil*)
FEDOCE	Colombian Federation of Educators
FEDESARROLLO	Foundation for Higher Education and Development
HLM	Hierarchical Linear Modeling
IAEP	International Assessment of Educational Progress
ICFES	Colombian Institute for the Advancement of Higher Education
IDANIS	Academic Diagnostic Instrument for Students Entering Secondary School (*Mexico*)
IDB	Interamerican Development Bank
IEA	International Association for the Evaluation of Educational Achievement
IEA	Aguascalientes Educational Institute (*Mexico*)
IES	Institute of Higher Education (*Brazil*)
LGB	Law of Guidelines and Bases for National Education (*Brazil*)
MECE	Program for the Improvement of Quality and Equity in Education (*Chile*)
MECE-RURAL	Program for the Improvement of Rural Schools (*Chile*)
MENA	Middle East and North Africa
MOE	Ministry of Education
NAEP	National Assessment of Educational Progress (*United States*)
NBPTS	National Board of Professional Teaching Standards (*United States*)
NCME	National Council for Measurement in Education (*United States*)
NPAs	National Plans of Action
NR	Regional Nuclei (*Brazil*)
OECD	Organisation for Economic Cooperation and Development
OTL	Opportunity-to-Learn
PAA	Academic Aptitude Test (*Chile*)
PAHO	Pan American Health Organization
PER	Performance Evaluation Program (*Chile*)

PLE	Primary Leaving Examinations (*Uganda*)
PME	Educational Advancement Projects (*Chile*)
PQE	Paraná Quality of Public Education Project (*Brazil*)
PREAL	Program for the Promotion of Educational Reform in Latin America and the Caribbean
PTA	Parent and Teacher Association
RAGS	Rescaled and Adjusted Gain Score model
REDUC	Latin American Educational Information and Documentation Network
SAEB	National System for the Assessment of Basic Education (*Brazil*)
SAT	Standard Attainment Task (*Great Britain*)
SAT	Scholastic Aptitude Test (*United States*)
SCC	School-Community Council (*United States*)
SEED-PR	Paraná Secretariat of Education (*Brazil*)
SEI	School Effectiveness Indices (*United States*)
SEP	Department of Public Education (*Mexico*)
SIMCE	National Educational Quality Assessment System (*Chile*)
SIP	School Improvement Plan (*United States*)
SMSO	Survey of Mathematics and Science Opportunity
SNED	National School Evaluation System (*Chile*)
SNP	National Testing Service (*Colombia*)
SNTE	National Teachers' Union (*Mexico*)
TAAS	Texas Assessment of Academic Skills (*United States*)
TAPs	Student Learning Workshops (*Chile*)
TIMSS	Third International Mathematics and Science Survey
UAA	Autonomous University of Aguascalientes (*Mexico*)
UNDIME	National Union of Municipal Directors of Education (*Brazil*)
UNESCO	United Nations Educational, Scientific, and Cultural Organization
UNFPA	United Nations Fund for Population Activities
UNICEF	United Nations Children's Fund
USAID	United States Agency for International Development
WHO	World Health Organization
ZEB	Basic Education Zone (*Mexico*)

INTRODUCTION

The value of education in shaping the history of man and society is incalculable, as testified to by philosophers through the ages, availing themselves of the most powerful analogies from each time and place to express what society hopes to gain from education and to change the course of prevailing practices. Aristotle, for example, used the prism of politics to delineate the functions of education in society. More recently, Piaget took his inspiration from the field of biology to represent the interactive processes taking place between the human organism and its environment that produce learning. Today, the metaphor of choice for reinventing education is the market metaphor, which emphasizes the dynamics of competition and market efficiency.

However, while these approaches add to our store of knowledge on human development and social progress and provide models for organizing education systems, they are inevitably circumscribed by their own basic assumptions and, as such, fail to exhaust all alternatives. Nor do they embody all aspects of the meaning of education in society. According to Kant, education is an art that takes innumerable generations to perfect and is man's greatest and most formidable problem. This natural limitation of our models of thinking and society's indefatigable concern with education are reflections of how teaching and learning are the human species' most distinctive feature. Moreover, society develops through learning mechanisms. Thus, it follows that one of society's ongoing tasks is learning about education. As a community of students and teachers, we've established evaluation mechanisms to assess our achievements and emerging needs, drawing, in particular, on modern-day social sciences in this process.

This book attempts to help spur the impressive educational reform efforts underway in Latin American and Caribbean nations by looking at the different dimensions of evaluation and at resulting options for framing public and private policy. Rather than simply design a model or present ready-made formulas, its main goal is to strengthen capabilities at the country level to continually reinvent education systems as one of the basic elements of the art of governance in our information-oriented society.

A great many people contributed to this publication, including Jim Hoxeng, from the U.S. Agency for International Development (USAID), who carefully read each paper with a critical eye, as well as Sarah Wright, who was involved in all facets of the planning and implementation of the studies serving as its foundation. Francy Hays, from the Academy for Educational Development, managed the project and provided invaluable guidance and leadership. John Engels and Tamara Mihalap, also with the Academy for Educational Development, provided production and editing assistance that greatly improved the quality of the publication. The authors and editors appreciate the work performed by the translators and revisers and, in particular, that of Ray Chesterfield from Juárez and Associates and of Ward Heneveld with the World Bank. This work was

initially presented at an international workshop on this issue sponsored by USAID's ABEL project in conjunction with the World Bank and PREAL (Program for the Promotion of Education Reform in Latin America and the Caribbean).

CHAPTER 1

THE CONTRIBUTION OF EVALUATION TO THE LEARNING OF NATIONS

Benjamín Álvarez H.

In few areas of social politics has it been possible to reach such a broad agreement throughout the hemisphere of the Americas as on the need to renew—and in most cases reform—national education systems. In the hope of resolving long-standing problems and ensuring themselves a definitive role in global society, the countries of Latin America and the Caribbean are putting into effect new educational policies that are notable for the fact that they:

- stress the results rather than the inputs of the educational process
- create opportunities for communities and civil society to form a commitment to education,
- provide greater autonomy to schools,
- promote the quality and efficiency of the systems,
- improve the professional competence of educators, and
- promote increased equality of opportunity.

As educational reforms continue to be implemented, the role played by government shifts from that of administrator to one of evaluator and policy maker. As more individuals and organizations participate actively in education, academic practices and student results become the subject of increased focus. The importance of the social learning promoted by the establishment of a culture of evaluation, together with appropriate evaluation procedures, will certainly increase, prompted by both political and practical need. This book introduces the reader to the problems involved in evaluating education in a climate of reform, the critical issues requiring policy decisions, and the undeniable contribution of the entire process to the learning of nations.

EVALUATION OF THE LEARNING OF NATIONS, ORGANIZATIONS, AND INDIVIDUALS

From earliest times, societies have alternately seen education as a way of reproducing values and knowledge and as a vehicle for achieving mobility and social change. Never

before, however, did they imagine that their destiny and their position in the world concert of nations would depend on their ability to learn, or that the "struggle to raise a nation's living standards is fought first and foremost in the classroom" (*The Economist*, 1997).

The oldest schools about which we have any information, such as those of the Sumerian civilization, had as their mission to give priority to the needs of selected groups of society, such as scribes and government functionaries. Things continued in this way for centuries. Today, however, education has ceased to be a matter of individual responsibility, motivation, and vocation, and has become instead a subject of public policy and a *sine qua non* for the survival of nations.

Trade, health, work, and human welfare are all dependent on the degree of participation by individuals and countries in the dense and intricate network of information and knowledge that envelops the entire world and is subject to an ongoing process of renovation. For the first time in history, human interaction can take place both simultaneously and globally, so that all peoples, to a greater or lesser extent but with no exceptions, find themselves involved in a learning venture unfettered by boundaries. The most optimistic visionaries of the social evolution of the recent past would be astounded by the opportunities created by cyberspace and the increasingly close relationship between knowledge and economics in terms of simultaneous diversity and the universal convergence of thought.[1] As the world becomes increasingly integrated as a result of that convergence, the one characteristic that will distinguish some regions from others will be the quality of their public institutions. Those achieving the greatest degree of success will perhaps be those with the most competent and efficient systems for supporting the collective interest, especially in terms of the production of new ideas (Romer, 1993). The development of such systems will in turn be dependent on the learning capacity and opportunities existing in a given society.

Learning to learn does not, then, constitute an exclusively individual goal; rather, it is also a national need. Nations, like individuals, evolve and change; they interact with their environment through processes of assimilation and adaptation. In a sense it can be said that they learn. But without mechanisms to rethink and revise their educational practices, to link their ideals and utopias to scientific knowledge and compare them with their actual achievements, national educational systems lose their bearings and sever their ties with their own identity. The problem no longer consists solely of developing specially designed contexts—such as schools, training centers, nonformal education programs—for the learning of all of the members of a society, but in addition of endowing such contexts with the capacity for constant assimilation and adaptation and the ability to seek optimum results. It is these results that, although intangible, form the basis for the well-being and progress of countries, for the productivity of businesses, and for the future of individuals (Porter, 1990; Reich, 1990). This is one of the reasons behind the increased political interest in the subject of evaluation.

This interest of nations, organizations, and individuals in improving the learning opportunities available to them is reflected in a number of convergent trends. The first,

and perhaps most public, expression of such trends is the recent proliferation of reforms being made to national education systems, based on civic participation and aimed at achieving top-quality results. The second is the indefatigable search for new models of organization from the productive sector that can be continuously adapted to the dynamics of the knowledge economy. The third is the increased need to accredit the learning currently being acquired by individuals in a variety of contexts throughout their lives.

Educational reform—with which most countries of the world expect to ring in the new millennium in response to the demands of the new economy—represents an opportunity for progress, provided that such renewed efforts are accompanied by an intelligent system of reflection, follow-up, readjustment, and reinvention. The history of educational reform suggests that there is no algorithm or magic formula to replace national and local capacity to learn, evaluate, and reform their educational systems. In fact, although the educational reforms of the current decade show great similarities in their ideological components and proposed strategies, their implementation and evolution follow very different patterns in individual countries. The most useful legacy of such reforms appears to be the strengthening of social values around the need to ensure a good education for all and the establishment of mechanisms to guarantee the continuous improvement of that education (Álvarez, 1997a).

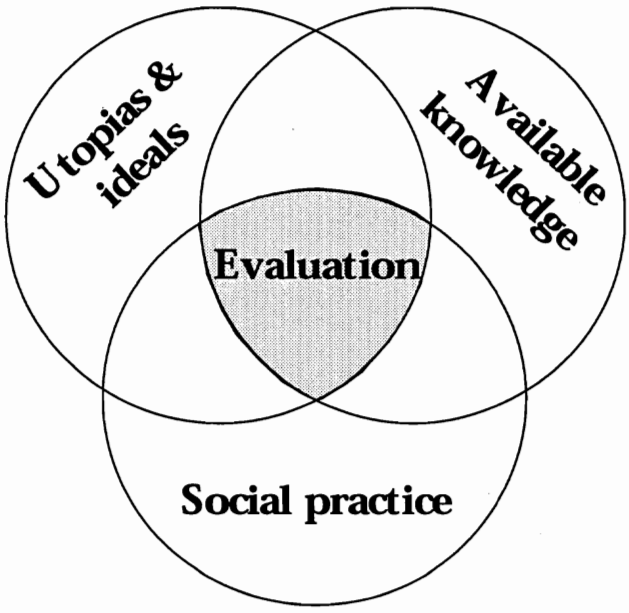
The commitment of governments to promoting the cognitive skills of their citizens cannot be limited to transient adjustments made to educational systems. Such a commitment involves decisions that simultaneously generate long-term effects on the patterns of distribution of both resources and responsibilities. In addition to resolving the inevitable dilemmas involving the nature and amount of the requisite investments, it is also necessary to define appropriate methods and to identify the institutions that will measure, value, develop, and disseminate information in this regard. Although we know little about knowledge as both an input and an outcome in the economy, information on the development, measurement, and use of human capital is becoming an essential point of reference for national policy. The predominant rhetoric attributes to education a value as an investment—as opposed to an expenditure—despite the fact that we do not have available sufficient theoretical and practical resources either to measure or to evaluate it. Economists, as well as accountants and educators, must aid in developing signs, indicators, and evidence for estimating, validating, and distributing information on human competence, so that such information can serve as the starting point for enhancing national social capital for the development and use of knowledge (Organisation for Economic Co-Operation and Development, 1996).

As businesses introduce increasingly greater knowledge and intelligence into their productive processes, their need for information that will enable them to identify and develop their human capital likewise increases. Just as governments do, businesses, private organizations, families, and individuals need to make decisions on an ongoing basis vis-à-vis their investments in education as a function of the economic, personal, and social benefits to be derived.

All of these concerns have led to a gradual broadening of the sphere of learning evaluation, which has grown from an initial interest in selecting individuals to occupy special positions or take advantage of subsequent educational opportunities to judging organizations and national learning systems. In Brazil, Colombia, Costa Rica, Chile, Mexico, and Venezuela, the first countries in Latin America to create national systems for testing scholastic achievement, the evaluation of academic learning has served for decades as a means for selecting individuals to take advantage of the scarce opportunities available for higher education. This parallels the situation in China, where the first tests of which we have any knowledge were developed some four thousand years ago. It was around such systems, the purpose of which was to identify candidates for public service at the county, provincial, and imperial levels, that higher education was organized (Wills and Lottich, 1961).[2] Nevertheless, new information needs are forcing those countries to expand the range of objectives of and approaches to educational evaluation, improve technical competence, and develop more effective communication strategies.

The most important challenge, however, does not consist merely of developing evaluation technologies and systems, but rather of strengthening the capacity of society as a whole to actively intervene in projects of collective interest—such as education—that determine its destiny; of promoting a culture that will favor evaluation, follow-up, and social responsibility; and of maintaining an infrastructure of knowledge (specialists, institutions, and networks) that will serve to transform closed and isolated testing systems into open projects of social learning.

Figure 1. The Dimensions of Evaluation



THE DIMENSIONS AND FUNCTIONS OF EVALUATION

Evaluation seeks to respond to the need for practical knowledge, the need to make decisions, and the need to reinvent particular objects or programs. However, the accumulation of the results of various evaluations begins to form a pillar of learning that inevitably transcends the immediate concerns of a given activity or program. Evaluation provides, in effect, an opportunity for the convergence of three dimensions: the utopias and ideals of a society, available knowledge, and social practice (Figure 1). Evaluation not only facilitates the bridging of policies and practices to desired goals, but also helps to draw the line that separates the ideal from the achievable.

These three dimensions or perspectives are present to a degree in every evaluation exercise, albeit in a state of permanent tension. Sometimes evaluations are guided primarily by the dynamics of knowledge, at other times by the needs of practice, and at still others by ideological, philosophical, or political precepts.

But in all cases, whether involving national indicators of human development, the measurement of organizational success, or the verification of personal progress, it is simply not possible to disregard the ingredients of knowledge, values, and utopias that are intermixed in differing proportions throughout the process of evaluation. This process enables countries, organizations, and individuals to compare themselves to others, to their own ideals, or to parameters based on scientific research. The evaluation of teachers, for example, combines criteria drawn from professional practice, current research, and the expectations of the society served by those teachers.

Because it is situated at the center of a dynamic of diverse interests, evaluation in education has served a number of purposes. In its most well-known and traditional version—school exams—evaluation has assumed, among others, a *symbolic function.* Evaluation marks the end of a cycle and provides credibility to educational processes by accrediting to society the achievement of specific learning or training objectives. On a broader scale, educational evaluation has demonstrated a *political function,* as it is linked to decisions regarding the destiny of individuals and to the survival of institutions and programs. In many cases, the information provided by the evaluation leads to the recasting of policies. This happens when, for example, comparisons between the indicators of the efficiency or effectiveness of the educational systems of several countries lead to the formulation of social policy goals or reforms in a particular country or state.

At other times, the opposite is true, i.e., social policy promotes and encourages the development of evaluation technology. Recently implemented policies of administrative decentralization have led in several cases to the development of social systems for providing follow-up and supervision of schools. Such is the case in the Brazilian states of Minas Gerais and Paraná, which have developed processes of self-evaluation for schools aimed at promoting their responsibility and improvement in a climate of increased autonomy. As a rule, the practice of evaluation is an exercise in politics, knowledge, and power, whether involving individuals, institutions or systems, small towns, or entire countries.

International competition provides an important thrust for national educational reform and the development of appropriate evaluation mechanisms and instruments. The late 1950s, for example, witnessed one of the most profound movements in educational renovation in the history of the United States as a result of its scientific confrontation with the Soviet Union. This movement introduced substantial changes to educational curricula, textbooks, and organization, and more than anything else, helped create a current of thought with regard to education that soon extended to other countries, particularly in Latin America.

The focus of this movement was to relate the goals and processes of education to the measurement and monitoring of results. Several of these works on evaluation—which went on to become classics in the field of education—stressed the importance of identifying the goals to be achieved through education and the way in which the scope of such objectives was to be determined. The objectives of education were organized around taxonomies or classifications identifying the various levels of thinking skills involved in the achievement of different objectives.[3] This systematic approach to teaching and learning had an enormous impact on educational planning throughout the continent and on the way in which results were conceived, at least in the central-level agencies responsible for curriculum design. Until recently, the academic curriculum in Paraguayan schools was designed on the basis of objectives expressed in terms of observable behavior, following the tradition initiated in the 1960s.

Private publishing companies contributed to the modernization of the teaching of science and mathematics, and some governments, despite the limitations of available evaluation technology and low levels of existing social demand, initiated or promoted the creation of scholastic achievement testing services that are still in operation today. This took place coincidentally with the development of public and private research centers created at the beginning of the 1970s for the purpose of studying educational problems and promoting innovations throughout the hemisphere.

Paradoxically, the end of the Cold War served to increase international competition rather than decrease it. The year 2000 has become the scenario of the battle to conquer markets. National education systems again face the stimulus of power and comparison with their peers. It is a case involving problems similar to those faced almost fifty years ago, but in a different context. Competition in a global economy based on innovation and knowledge is translated into an increased interest in achieving an education for all of the citizens of a country that meets higher standards of quality. Standards are defined as a goal and as a measure of the degree of progress toward the achievement of that goal. Ravitch (1995) mentions three interrelated types of standards: standards of content or curriculum (that which should be taught), standards of scholastic performance (levels of achievement), and standards of learning opportunities (availability of programs and resources). The movement in favor of the standards of education has, in addition, an international horizon, particularly in the fields of science and mathematics, which have been the focal point of most comparisons of academic achievement among countries.

This climate of global competition makes the educational systems of the countries of the Americas increasingly aware of the three fundamental gaps that they face: the disparity of human resources existing among the countries of the region and among the various regions of the world, the persistent—and in some cases increasing—inequalities observed within each country, and the great distance separating the advance of knowledge from the learning opportunities being offered to children and youth. Evaluation makes it possible to acquire a better understanding of these gaps and of available policy and program alternatives.

This consideration leads us to affirm that evaluation also fulfills a *function of knowledge.* The professional practice of evaluation, particularly in the United States, has stressed its purpose as a systematic decision-making mechanism with regard to the value or merit of a particular social object or program, with a diminished interest in its contribution to the knowledge of subjects such as the efficiency of social interventions and the learning of organizations. "Practical knowledge" as defined by Aristotle—as opposed to "speculative knowledge" (i.e., knowledge for its own sake)—is a most useful category for describing the nature of the knowledge generated by evaluation. This type of knowledge seeks useful results, and its purpose is to guide human action. It involves, according to the philosopher, the use of innovative and creative faculties, the ability to organize, and the desire to act. This strategic knowledge of the functioning of social interventions and educational institutions is one of the most useful products of evaluation. It is this knowledge that serves as a basis for the *function of improvement* of the objects of evaluation, which has been emphasized as its primary characteristic, in order to distinguish it from research, the purpose of which would be to confirm theories.

No less important than the above is the *function of developing capacity* that evaluation entities can promote in a country. For example, the programs on which educational reforms are based are generally characterized by their intense activity and their brief duration, especially when financed by international loans or grants. Unless a country or state develops a monitoring and follow-up infrastructure that enables it to rethink and revise the changes being implemented, it will find itself in a situation of ongoing uncertainty. This infrastructure consists primarily of a critical mass of analysts, an institutional base, and networks for interaction and information—in other words, of human resources trained in the tasks of evaluation, responsible public or private organizations, and efficient channels of communication and participation.

EVOLUTION OF THE PRACTICE OF EDUCATIONAL EVALUATION IN THE HEMISPHERE

Two principal tendencies underlie today's educational evaluation in the Americas. The first had its origin in experimental psychology and psychometrics, in which theories and measuring instruments were designed to assess intelligence, skills, performance, and academic learning. The second tendency was a social science movement that grew in response to the growing needs for orientation and guidance for intervention programs, policies, and social investments.

The technology and practice of the evaluation of a variety of individual and learning characteristics have been notably enriched since the first tests—developed by Binet—were used in France for the purpose of identifying children with mental disabilities. Buyse, a disciple of Binet, and other European professors introduced the fundamental ideas of the nascent field of psychometrics to Colombia during the 1930s (Restrepo, 1995). Since then, Colombia, along with other countries such as Brazil and Chile, has become a pioneer in this field in Latin America. But it has been in the United States that demand for testing services has been the most intense and that the increase in availability has been greatest.

At the University of Costa Rica, a testing service was organized in 1986 with the goal of improving the educational system (as opposed to aiding universities in their student admission procedures, which was the predominant objective of the testing services existing at that time) (Esquivel, 1996). Chile did likewise with its System for Measuring the Quality of Education in 1988. Several countries of the region have opted for a similar scheme, with the result that the new testing services developed during the current decade are aimed at improving the quality of the educational system, unlike those developed and used by their predecessors. Table 1 presents a list of countries that initially developed selective processes for measuring academic achievement. It also lists a second generation of systems (almost all developed subsequent to 1985) that were inspired by the analogy of the efficient business. These systems represented an attempt to serve as both a reference and a stimulus for improving the quality of education. A third generation of broader evaluation approaches—one that uses multiple analogies and attempts to affect the educational system in several critical areas, in addition to its final results in terms of academic achievement—is now beginning to emerge.

The importance of educational evaluation has in addition been officially recognized in several countries. It is mentioned in Colombia's 1994 General Law of Education (Restrepo, 1995) and in a Constitutional Organic Law passed in Chile during the military regime (Rodríguez, 1996). Brazil's Law of National Education Standards and Guidelines (1996) includes among the tasks falling to national education entities: "to guarantee the existence of a national process for evaluating scholastic performance in primary, middle and higher education in collaboration with teaching systems, by providing for the definition of priorities and improvements in the quality of teaching."

Some national constitutions are beginning to reflect the effects of the changing role of the state with regard to education and the new emphasis on guaranteeing not only access to but also the quality of that education. Article 16 of Peru's National Constitution states that:

> Both the educational system and the educational regime are decentralized. The State coordinates educational policy. It formulates the general guidelines for plans of study as well as the minimum requirements for the organization of educational institutions. It supervises compliance with those requirements and the quality of the education provided.

State educational laws in larger countries are also beginning to reflect the concern for ensuring that educational activities produce the expected results, as illustrated by the recently approved Law of Education of the State of Aguascalientes (1997) in Mexico, which authorizes the Education Institute to create a State Educational Evaluation System.

Table 1. National Scholastic Achievement Testing Systems

Countries	Purposes of selection and promotion	Quality-oriented
Argentina		✔
Brazil	✔	✔
Chile	✔	✔
Colombia	✔	✔
Costa Rica		✔
Cuba	✔	
Dominican Republic		✔
El Salvador		✔
Guyana	✔	
Jamaica	✔	
Mexico	✔	✔
Venezuela	✔	✔

Sources: Interamerican Development Bank (1997); Álvarez (1995)

Research on education flourished throughout the hemisphere with the dramatic expansion of educational systems in the 1960s and 1970s. However, with the exception of Canada and the United States, the development of a methodology and a systematic approach to the evaluation of programs and institutions began only recently. In the United States, the large social intervention programs promoted by the federal government in the mid-1960s to aid socially disadvantaged children and youth required empirical evidence of their effect. This requirement strengthened the development of a movement toward evaluation that led to the creation of, and experimentation with, working models and the configuration of an academic community (Worthen and Sanders, 1987). This movement was preceded, however, by an initial stage in the 1940s and 1950s that laid the conceptual bases.

This community has attempted to identify principles to guide the professional practice of evaluation, such as the standards for the development and application of psychological and educational tests in 1966, the standards for the evaluation of educational programs (The Joint Committee on Standards for Educational Evaluation, 1994), and the standards for the evaluation of personnel in 1988.

In Latin America, educational research has been exceedingly rich in its theoretical conceptions and in the design of innovations closely related to the social problems affecting the region, particularly the gaps in development and the alienation of large majorities of people. It has been less productive, however, in terms of the measurement and analysis of the results and the quality of education, which is the declared purpose of most educational revitalization programs being implemented in virtually all of the countries of the region with support from international banks and agencies.

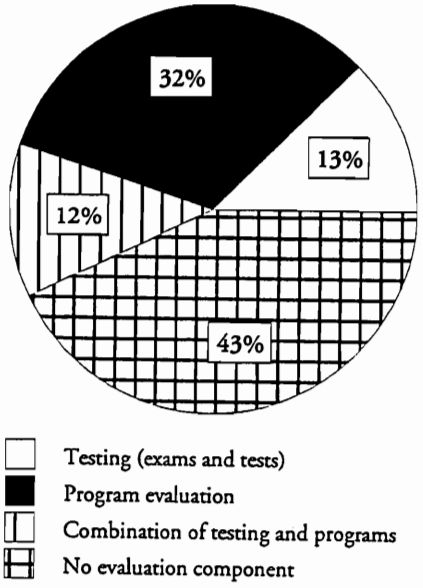
Information gathered in 1996 on a sample of sixty educational programs, either in operation or in the process of being approved, financed by the World Bank, the Interamerican Development Bank (IDB), and the United States Agency for International Development (USAID) (Álvarez, 1997), shows that a large majority of the programs (67 percent) include an evaluation component. Thirteen percent of the programs explicitly involve exams or tests of scholastic performance (Figure 2). However, this renewed interest in learning the effects of the school system contrasts with the notable deficiencies of educational information systems in Latin America and the Caribbean. This is one of the problems that the Secretariat Pro Tempore of the Americas has identified in its collective effort to follow up on human development goals in the hemisphere.

Notwithstanding the growth and expansion of educational evaluation activities in the Americas, there still exist significant limitations that prevent both the scientific community and the general public concerned with education from influencing policy decisions. There has yet to be developed to a sufficient degree a "culture" of evaluation whose various components make sense and that facilitates a coming together of the various groups making up the national educational community. Nevertheless, there is available an abundance of experiences, knowledge, and instruments that can provide top-quality assistance to countries and academic systems alike in allocating resources and improving opportunities for human learning and development.

THE CRITICAL AREAS OF EDUCATIONAL EVALUATION

The fragility characterizing the relationship between the scientific community, families and schools, and decision-making institutions is of particular concern in Latin America and the Caribbean for two primary reasons. First, despite the predominant political rhetoric, it has not been possible to establish an attitude of collective accountability with regard to public education or efficient mechanisms for monitoring school system performance. Education as a project of society as a whole is still a distant ideal. It might be possible to extend to almost all of the countries of the region the argument made in Brazil that the greatest obstacle to educational reform is the predominance of private interests over public goals (Plank et al., 1994). And second, new social policies tend to assign new responsibilities to communities lacking resources and with little "convocation capability," or sufficient influence to involve others in their educational effort.

Figure 2. Educational Evaluation in International Projects



Source: Academy for Educational Development (1996)

Added to the above weakness in terms of the ability of civil society to participate in national educational decisions are the limitations of those very systems in controlling the educational enterprise. A recent analysis conducted by the Interamerican Development Bank (1997) of sector studies of primary education carried out by the World Bank and IDB in fifteen countries of Latin America and the Caribbean between 1991 and 1995 revealed that eight countries failed to conduct an evaluation of the results obtained in terms of student learning; six had no operational information systems in place; ten had inadequalte systems for supervising teacher performance; and in six the Ministry of Education was shown to have limited planning, research, and evaluation capacity (Table 2). These findings lead to the conclusion that there are two needs, to strengthen national capacity and to satisfy the urgent need to evaluate current reform programs.

There are four closely interrelated areas of information and evaluation that are critically important in terms of social policy and that a country or state must define for the education sector:

- the results obtained from the sum total of the educational efforts and enterprises of society,
- the significance and efficiency of the educational process,
- the professional capacity of the actors directing the process: teachers and administrators, and
- the immediate learning context, which in most cases is the school.

Follow-up of the results of education in the national environment

Formal education has always had a way of accrediting its effects on an individual scale by means of examinations, grades, and diplomas. But as the educational history of nations has progressed and as more people study increasingly complex subject matters for longer periods of time, the need for aggregate information on education at the state, national, and global levels has increased. The information coming from the educational sector is in turn one of the ingredients in the process of analyzing the progress of countries and designing social policies. Although educational information began to be collected as far back as the era of the rapid expansion of educational systems in Latin America and the Caribbean (1960s and 1970s), there is a widespread consensus that such information is no longer sufficient as an input for determining the results of national educational efforts, nor for undertaking actions aimed at improving those results.

Table 2. Evaluation Problems Identified in International Projects in Fifteen Countries in Latin America, 1991-1995

	Inadequate teacher supervision	No evaluation of learning	Limited capability in Ministry of Education	No information systems	No community participation	Concentration of decision-making authority	Inadequate teacher incentive
Argentina				✔		✔	✔
Uruguay	✔	✔			✔		
Chile	✔		✔	✔			
Bolivia		✔			✔	✔	
Brazil	✔	✔		✔			✔
Peru			✔	✔			
Mexico	✔	✔			✔		
Venezuela	✔	✔			✔		
Paraguay	✔	✔					
El Salvador	✔	✔	✔				✔
Guatemala			✔			✔	
Honduras	✔	✔			✔		✔
Costa Rica	✔		✔	✔		✔	
Jamaica							
Barbados			✔	✔			
Number of countries	10	8	6	6	5	4	4
Percentage of countries	67	53	40	40	33	27	27

Source: Interamerican Development Bank (1997)

Available information on the results of education consists, in most cases, of statistics of questionable reliability, involving indicators of formal aspects of educational systems (access, retention, etc.) and data obtained in some countries through tests and exams. The English-speaking countries of the Caribbean, following the European tradition, developed a system of national public exams, while some of the countries of Central and South America (Table 1) generated information by the application of standardized tests or some type of measurement of scholastic achievement based on samples of students enrolled at the elementary school level. However, these resources are insufficient for responding to the needs created by the new emphasis on the quality of the results of national educational systems. In this regard, educational policy in the late 1990s in all of the countries of the Americas faces questions such as the following:

- What are the objectives, the clientele, and the usefulness of national evaluation systems?
- What information should be included in national systems for evaluating learning?
- Is the setting of goals vis-à-vis the evaluation of the results of education a matter of national policy or local decision making?
- What is the contribution of goals and standards to educational improvement?
- Is the formulation of national standards for scholastic achievement desirable, advisable, and feasible?

The dilemmas facing countries endowed with low levels of resources are even more pressing. A fundamental task for these countries is to estimate whether the benefits of investments in systems for following up on the results of education are greater than those that could be achieved by improving educational inputs or organization.

Evaluation of the schooling process

Knowledge of the results of education is an essential, but not sufficient, element in the evaluation of educational systems. The application of tests, measurements, or verifications of scholastic achievement does not in itself raise the quality of service. In the absence of a constant flow of information on what is being taught and how it is being taught, it is not possible to explain the results nor to formulate guidelines with regard to the desirable practices and professional requirements of teaching personnel.

In today's climate of reform and grandiose formulations, it is appropriate to recall the lessons of history. A goodly portion of the curricular transformations of the past failed to take root fully because they were unable to affect the daily life of the microcosm of the school and classroom. Decisions involving the policy of school curriculum follow-up are not exempt from tension and dilemma. In fact, if it is accepted that the goals of education must be formulated by the national government for the purpose of obtaining high-quality results, what possibility for participation in the core of the educational process remains for autonomous scholastic communities, educational professionals, and the many other imaginable actors in a democratic society?

- What type of monitoring would enable a balance to be struck between the desirable national curriculum, the curriculum planned in each school and town, and the actual curriculum being implemented? Is this the same as the existing relationship between what to teach, as determined at the central level, and how to teach, as carried out locally?
- What policies would ensure consistency between the school curriculum, systems of tests and examinations, and academic texts and materials?
- Should educational reform be guided by curricular innovation, by the restructuring of school organizations, or by standard-setting and evaluation?

Evaluation of school organizations

The concern for the efficiency of school organizations has doubtless been prompted by what is happening in the field of contemporary productive organization. But such concern is not entirely new on the continent. A little more than a century ago, the British Crown introduced, initially in Barbados and subsequently in other countries of the Caribbean, a system of results-based payments to schools that was built around examinations designed "to test the character of the teaching being given and the overall administration of primary schools" (King, 1995). The inspectors of that time prepared "standards" and scores and established efficiency ratings to be used as an input in setting financing policy. The questions that we pose today are perhaps not altogether different. For example:

- Is it fair to evaluate a school on the basis of the results obtained by its students as compared to other students from other schools?
- How much can a school learn from productive organizations with regard to school management and evaluation without losing its character and without losing sight of its ultimate objective?
- How would an effective school be defined?
- What organizational factors and conditions are suggested by empirical evidence as being associated with the success of a school?

Evaluation of the teaching profession

Although teaching may not currently enjoy the same halo of dignity that surrounded it throughout history, its success, as measured in terms of its effect on the life of students, is perhaps more necessary now than in the past in order to ensure human advancement. But the qualitative and quantitative importance of the teaching profession contrasts with the lack of interest shown by governments, society as a whole, and teachers themselves in the professionalization of their activities—a professionalization that involves social responsibilities and, obviously, mechanisms for teacher selection, promotion, and encouragement in addition to the evaluation of teacher performance.

Problems of conceptualization regarding the proper definition of the basic ingredients of effective teaching and the low levels of interest shown by researchers, political

problems such as the prevalence of weak education ministries confronting strong teacher unions, and administrative problems such as the centralization or nonexistence of teacher incentives have hampered the development of mechanisms to improve the level of professional competence of teachers. Evaluation is, in effect, useful not only for administrators and parents, but also for teachers themselves. However, in this field, policy faces difficult-to-resolve issues and dilemmas, among which are the following:

- What purpose is to be served by teacher evaluation?
- What criteria should be used to evaluate teacher performance?
- Who is responsible for teacher evaluation?
- At what point during the professional teaching career is evaluation most useful?
- What does our experience with existing evaluation tests show us?

This book addresses the four critical areas of social policy and evaluation—the results of formal education, educational processes, teachers, and school organizations—by combining information drawn from three sources: international research, the experience of countries having achieved the greatest degree of success in the field of evaluation, and the implementation of recent innovations which, in the field of evaluation, have been motivated by the educational reforms of the 1990s taking place in a number of countries of the hemisphere. Its purpose is to contribute to the learning of nations through the identification and analysis of the dilemmas and opportunities occurring on both the individual and collective planes in countries, states, and families.

The book is organized into four parts. The first part, "National Systems," begins with a chapter on monitoring the results of education in the national environment, which provides an overall scenario for the subject of following up on the performance of national education systems. For most countries, and particularly those with low levels of resource endowment, the dilemma involves deciding whether to invest in the evaluation of scholastic achievement or in basic inputs for the proper functioning of the system. The first option leads to new dilemmas: What types of follow-up on the effects of education are both timely and useable? What institutional, technical, and financial resources are required? Although interest in the subject of the aggregate results of education is relatively new, several countries possess considerable experience in evaluating scholastic achievement by following a variety of schemes that have served as points of reference for a consideration of the overall products of educational systems. This chapter integrates the information provided by international experience with the analysis of the results of social research and with the practical problems identified by the policy making process in developing national follow-up systems for school learning.

The second chapter presents an analysis of the evaluation of the processes that supposedly lead to the results achieved by educational systems. Its focal point is the follow-up and monitoring of the school curriculum. The school curriculum is, in effect, a central instrument of educational policy. This chapter illustrates the relationship between evaluation and policy in the context of recent international analyses and comparisons of scholastic learning. The first part of the book concludes with a discussion of the case of

the international monitoring of the goals of human development initiated by the countries of the Americas as a result of the 1994 Nariño Pact.

The second part, "The Lessons of History," attempts to recount the experience of the two countries of Latin America with the greatest tradition of evaluation by means of an identification and discussion of the social effects that educational evaluation systems have had in recent decades. The essential questions discussed in these chapters are: What has been the significance—for families and students, for social policy, and for educational organizations in the countries—of the operation of a system for periodically evaluating scholastic achievement and what lessons can be derived from this experience to benefit future educational policy?

The third part of this book, "Teacher Evaluation and Professionalism," discusses the state of the art in teacher evaluation and reviews the practice of teacher evaluation in five countries in Latin America. In addition to the above-mentioned theoretical difficulties, teacher evaluation is subject to political pressures from educator unions and professional groups and associations and is limited in many countries by the lack of an appropriate definition of its purposes. This situation is compounded by the fact that the new dynamic created by the processes of decentralization and scholastic autonomy is forcing local governments and scholastic communities to seek mechanisms for controlling and developing the teaching profession. The governments of the region find themselves faced with a dilemma similar to that encountered by Pliny the Younger who, when wishing to establish a school in his town and opting to finance only a portion of its cost, based his decision on the following rationale:

> I would promise the whole amount were I not afraid that some day my gift might be abused for someone's selfish purposes, as I see happen in many places where teachers' salaries are paid from the public fund. There is only one remedy to meet this evil: if the appointment of teachers is left entirely to the parents, and they are conscientious about making a wise choice through their obligation to contribute to the cost.

The problem of financing educational systems, the social responsibility of the directors of such systems, the professional development of teachers, and their influence on the quality of the results of education are united by available evaluation institutions, which as a rule are timid, insufficient, and inoperative.

The primary task of education is to provide an appropriate environment and a stimulus for learning. The school is a prototype of this context. It is there that the expectations, actors, and resources for learning converge. The chapter which begins the fourth part of this book, "The Evaluation of the Organization of Education," provides a framework for placing policy options for evaluating schools within the context of national or state reform programs. It describes such options within the academic tradition of evaluation in general and of research on the effectiveness of schools in particular, and presents an alternative model, the aggregate value model, with which the state of Texas has experimented. This model includes not only the evaluation of schools in particular but also

the evaluation of school systems in general. The problem of the efficiency of the educational enterprise in Latin America has been highlighted in recent international analyses and comparisons (Interamerican Development Bank, 1996).

The chapters that follow describe two of the innovations being tested in several countries (Paraná state in Brazil and Aguascalientes state in Mexico) with a view toward accompanying and evaluating the process of educational reform within the very enclave in which it takes place. The document concludes with a brief analysis of emerging issues and of the policy alternatives available to the countries.

CRITERIA FOR AN EVALUATION POLICY AGENDA

The evaluation of education is an essential component of the learning mechanism of nations. It makes it possible not only to determine the progress, setbacks, opportunities, and achievements of their peoples, and to make more accurate judgments, but also to move toward the identification of collective ideals. It is only through the reflection stimulated by evaluation that it is possible to transcend the predominant rhetoric that, in a climate of reform such as exists at present, can prevent the exploration of new alternatives. As an instrument of social learning, evaluation must fulfill requirements generated by three types of criteria:

- significance and social usefulness,
- orientation toward the future, and
- precision.

In effect, evaluation is not exclusively technical in nature, because in the absence of a broad social base and a national community with a capability and an interest that translates into political will, evaluation institutions or programs are inoperative and socially nonexistent. Moreover, the collaboration and negotiation involved in the evaluation exercise constitute one of a society's most useful channels for interaction with regard to the development of its own identity. Accordingly, a national evaluation policy should respond to the needs for the integration and participation of the various groups of society responsible for, and concerned with, education. In addition, such a policy requires a comprehensive view of the areas making up the critical subject matter involved in the decision making process.

Although the immediate objects of evaluations are the footprints of the past and the information of the present, they become meaningful and contribute to human and social development to the extent that their purpose is a particular action in the future, which is their field of application and existence. The second group of criteria for an evaluation policy agenda thus includes those criteria that contribute to ensuring that evaluation activities will serve the future of educational systems and encourage their capacity for change.

The technical quality of the evaluation is an indispensable prerequisite for its reliability and use. This quality is not limited, but rather expands as a result of the incorporation

of alternative methodologies and new disciplines and metaphors. In the romance languages, the term "to evaluate" has the original meaning of "to weigh" and evokes the image of the scales that make it possible to make judgments with knowledge and precision. The symbol of precision, for the ancient Egyptians, was the feather (the goddess Mayet, representative of justice, truth, and order in the world) that served as a weight in the pan of the scales used to weigh souls.[4] Precision is, in effect, the characteristic that makes it possible to issue a weighted judgment and generates true practical knowledge on which the learning of nations is built.

NOTES

[1] T. de Chardin, for example, envisioned a progressive humanization and expansion of the network of thought that would be at once diversified and converging.

[2] The complex system of testing in China served as a screening tool from the county level all the way up to the capital city. Candidates that were able to pass national tests were entitled to hold public office. (Wills, E. and Lottich, V., 1961. *The Foundations of Modern Education.* New York: Holt, Rinehart and Winston).

[3] The taxonomy of the objectives of education proposed by a series of experts headed by Benjamin Bloom of the University of Chicago in 1956 is quite representative of this movement.

[4] Cited by I. Calvino (1994). *Seis Propuestas para el Próximo Milenio.* Madrid: Ediciones Siruela.

REFERENCES

Álvarez, B. (1995). "Equity and Selective Access to Education in Latin America." In T. Kellaghan (ed). *Admissions to Higher Education: Issues and Practice.* Princeton: International Association for Educational Assessment.

Álvarez, B. (1997a). "Life Cycle and Legacy of Educational Reforms in Latin America and the Caribbean." Presented at the annual meeting of the Comparative International Education Society, Mexico City, March 19-23.

Álvarez, B. (1997b). "Naturaleza y contexto de las reformas educativas de final del siglo." In B. Álvarez and M. Ruiz-Casares (eds). *Senderos de Cambio: Génesis y Ejecución de las Reformas Educativas en América Latina y el Caribe.* Washington, DC: Academy for Educational Development.

Calvino, I. (1994). *Seis Propuestas para el Próximo Milenio.* Madrid: Ediciones Siruela.

Coulson, A. (1996). "Markets versus Monopolies in Education: The Historical Evidence." *Education Policy Analysis Archives,* 4: 9.

Esquivel, J. (1996). "Medición de Logros del Aprendizaje y el Empleo de los Resultados en Costa Rica." Presented at the International Seminar on Educational Evaluation and Standards in Latin America: Realities and Challenges. PREAL, Río de Janeiro, December 4-5.

Interamerican Development Bank. (1997). *Progreso Económico y Social en América Latina: Informe 1996.* Washington, DC.

King, R. (1995). "Education in the British Caribbean: The Legacy of the 19th Century." *La Educación. Revista Interamericana de Desarrollo Educativo,* 34, 121:243-260.

OECD. (1996). *Measuring What People Know: Human Capital Accounting for the Knowledge Economy.* Paris.

Plank, D., Sobrinho, A. and Xavier, A. (1994). "Obstacles to Educational Reform in Brazil." *La Educación. Revista Interamericana de Desarrollo Educativo,* 33, 117:75-95.

Porter, M. (1990). *The Competitive Advantage of Nations.* Boston: Harvard University Press.

Ravitch, D. (1995). *National Standards in American Education: A Citizen's Guide.* Washington, DC: Brookings Institution.

Reich, R. (1990). *The Work of Nations: Preparing Ourselves for the 21st Century Capitalism.* New York: Vintage Books.

Restrepo, G. (1995). "Exámenes nacionales universitarios de ingreso y de egreso. Su relación con el sistema nacional de la calidad educativa". In *Misión nacional para la modernización de la universidad pública.* Bogotá: Editorial Presencia.

Rodríguez, C. (1996). "Sistema de medición de la calidad de la educación. Características y usos de los resultados de la evaluación para mejorar la calidad." Presented at the international seminar on educational evaluation and standards in Latin America. Realities and challenges. PREAL, Río de Janeiro, December 4-5.

Romer, P. (1993). "Two Strategies for Economic Development. Using Ideas and Producing Ideas." Proceedings of the World Bank Conference on Development Economics.

The Economist. (1997). "Education and the Wealth of Nations," March 29.

The Joint Committee on Standards for Educational Evaluation. (1994). *The Program Evaluation Standards: How to Assess Evaluations of Educational Programs.* 2nd edition, London: Sage Publications.

Wills, E. and Lottich, V. (1961). *The Foundations of Modern Education.* New York: Holt, Rinehart and Winston.

Worthen, B. and Sanders, J. (1987). *Educational Evaluation: Alternative Approaches and Practical Guidelines.* New York: Longman.

Section I
NATIONAL SYSTEMS

CHAPTER 2

MONITORING NATIONAL EDUCATIONAL PERFORMANCE

Thomas Kellaghan

The introduction of the social policy reforms that herald the dawn of the twenty-first century has increased considerably the pressure on educational systems to achieve results that meet the expectations of nations. Politicians, academics, and parents for whom channels for participation in education are opening require information on the performance, not only of the students as individuals, but also of educational systems as a whole. Policy dilemmas are inevitable when determining where efforts are to be focused and what forces are to guide change. This chapter responds to the first group of concerns described in the introduction, i.e., those related to the results of the sum total of the educational efforts and enterprises of society. The question is simple but far-reaching: What do students learn at school? This question makes it possible to address the problem of evaluating national educational systems from the multiple perspectives of the researcher, administrator, and politician. It also makes it possible to organize the elements of the debate into the categories of international experience, recent research, policy options with regard to the scenario, and procedures involved in the creation and operation of a national evaluation system.

Most government departments have for a long time routinely collected and published statistics that indicate how their education systems are working and developing. Statistics are usually provided on a variety of inputs—such as school numbers, facilities, and student enrollments—and efficiency indices, such as student-teacher ratios and rates of repetition, drop-out, and cohort completion. But despite an obvious interest in what education achieves and the substantial investments of effort and finance in its provision, few systems in either industrial or developing countries have, until recently, systematically collected and made available information on the outcomes of education.

Information on outcomes—what students learn at school—has traditionally been associated with the assessment of individuals and has been used for a variety of purposes, including the monitoring of student progress, the diagnosis of problem learning

areas, the motivation of students, and the guidance of remedial action. All of these uses may be regarded as formative; that is, the information derived from assessment is used to interact with and improve the students' learning. Assessment, of course, may also be summative when it provides a statement of the students' knowledge and skills at the end of a learning program.

In recent years, assessment has been extended to school systems, where it has been used for purposes analogous to those used with individual students. It too can be summative, describing outcomes without reference to treatment, or formative, in which case it is used for diagnosis, monitoring, motivating, and guiding remedial action. When assessment is used at the system level, we may follow the usage of the United States and refer to it as national assessment, though such assessment is not confined to the United States. The United States does, however, provide an outstanding example of a national assessment system—the National Assessment of Educational Progress (NAEP)—which held its first survey in 1969. In Great Britain, the Assessment of Performance Unit began operation in 1975. The British system has since been superseded by a more elaborate one, which has been the source of much controversy. There is also a national assessment system in France and in many other countries across the world: in several European countries, Canada, Latin America (Costa Rica, Chile, Colombia, Venezuela), Africa (e.g., Egypt, Mali, Mauritius, Morocco), and Asia (e.g., China, Hong Kong, Thailand) (see, for example, Chinapah, 1996).

The focus on school outcomes gives rise to a relatively simple question: What are students learning at school? The answer, however, is quite complex and raises a variety of issues. First, given that schooling has a variety of goals, what outputs are to be taken as representing the effects of schooling, who is to decide on them, and by what procedure? Second, what methods can be used to measure the outcomes of the education system? Third, having decided on a method of assessment, how can we be confident that the information it yields adequately represents what is achieved in the education system (or in a part of it)? Fourth, what is the best way to report information on outcomes? And finally, if we wish to engage in formative action, how can findings on the achievements of the education system be turned into action?

The purpose of this paper is to address these issues, all of which are relevant to policy makers who might be considering the introduction of a system for assessing and monitoring the outcomes of their education system. The nature and rationale of national assessments will be considered, as well as their many components, the variation in how they are constructed in different countries, and their cost.

NATIONAL ASSESSMENTS AND THEIR RATIONALE

A national assessment may be formally defined as a procedure to measure the learning outcomes of an education system. I will consider the purposes of and uses to which a national assessment may be put in greater detail below, but at this point the general

purpose of a national assessment may be said to be to provide those involved in policy and decision making in education with information on students' achievements that is designed to improve decisions and suggest cost-effective interventions to improve learning.

Many factors have led to a situation where countries are concerned about learning outcomes and are prepared to invest resources to find out what knowledge, skills, and competencies students are acquiring at various stages of the education system. One factor is that we can no longer accept that inputs can be used as reliable proxies for outputs. That is, we cannot assume that because a child has been in school for four, five, or six years that he or she has, in the words of the World Declaration on Education for All, actually acquired "useful knowledge, reasoning ability, skills, and values" (UNESCO, 1990, par 4). A second reason for developing national assessments arises from the fact that economic and technological changes are demanding higher levels of knowledge and skills among school leavers and that some system of monitoring is required to inform policy makers and education managers about the extent to which this is happening.

The need to improve students' achievements is underlined by a consideration of the effect of increasing free trade and competitiveness between nations in economic activity. In this context, belief has been rekindled in the role of education in enhancing a nation's supply of "human capital." It is argued that if a country does not have an effective education system, it will not have the competent, productive, and competitive work force necessary to maintain and improve economic performance and to increase prosperity (Guthrie, 1991). Thus, the outcomes of the education system should be systematically monitored so that governments can be assured that they are satisfactory to meet their economic goals.

A further reason for the development of national assessments is that they are considered necessary to address concerns that the efficiency of the education system needs to be improved. This arises from the fact that in many countries, governments are faced with the problem of dealing with expanding enrollments while at the same time being asked to improve the quality of education without increasing expenditure. This requires increasing the efficiency of education systems. However, to obtain evidence on whether or not this efficiency is being achieved, it is necessary to have information not only on inputs, but also on outputs (Lockheed and Hanushek, 1988).

While it may be difficult for any country, whether industrialized or developing, to resist the pressures created by these considerations, developing countries should pause and consider seriously whether or not they should initiate a national assessment. The question to be answered is: Should a country use its limited resources to emulate the practice of richer countries in monitoring educational outcomes, or should it apply those resources to improving educational inputs, such as school buildings, teacher education, textbooks, laboratories, or other facilities?

INDICATORS

The kind of assessment task that is used for a national assessment may not differ very much from the kind of task that a student would take in an assessment to monitor his or her individual scholastic progress. Thus, if one were to watch a student complete an assessment task, one might not be able to say whether the task was part of an individual or national assessment. However, once the student has completed the assessment, the way in which results are used differentiates an individual assessment from a national assessment.

In a national assessment, data on the performances of students are aggregated, usually for all students in the country or for a representative group of them, and perhaps also for varying groups (e.g., girls, boys, urban students, rural students). When aggregated in this way, the performances are treated as *system outcomes* or *indicators*, to use another term that has its origins in economics.

For data to be regarded as indicators, they should exhibit certain characteristics:

- An indicator is quantifiable: it represents some aspect of the education system in numerical form.
- A particular value of an indicator applies to only one point or period in time.
- A statistic qualifies as an indicator only when there is a standard or criterion against which it can be judged. The standard may involve a norm-referenced (synchronic) comparison between different jurisdictions; a self-referenced (diachronic) comparison with indicator values obtained at different points in time for the same education system; or a criterion-referenced comparison with an ideal or planned objective.
- An indicator provides information about aspects of the education system that policy makers, practitioners, or the public regard as important. Sometimes it is easy to obtain consensus among interested parties on what is important; other times it is not.
- An indicator is realistic in the sense that it is based on information collected with due regard to financial and other constraints.
- An indicator describes conditions amenable to improvement.
- Information for indicators is collected frequently enough to allow change to be monitored.
- An indicator allows an examination of distributions among subpopulations of interest (for example, age, gender, or socioeconomic group) (Greaney and Kellaghan, 1996).

The selection of indicators to represent the status of the education system should be based on a model, which may be explicit or implicit, of how the education system works (Burnstein, Oakes, and Guiton, 1992). Further, the set of indicators incorporated in the model should reflect the multifaceted nature of education in all its complexity (Bottani and Tuijnman, 1994) as well as being comprehensive enough to describe the important dimensions of the system. The model, in turn, should provide a context for interpreting what the indicators mean, describing how they relate to other aspects of the

education system (and perhaps to other social and economic systems), and suggest how they are likely to respond to various kinds of manipulation.

The model of the education system on which current systems of indicators in most countries are built is one that dominates research in school effectiveness and involves a consideration of inputs, processes, and outputs. *Inputs* are the resources available to the system, e.g., buildings, books, the number and quality of teachers, and such educationally relevant background characteristics of students as the socioeconomic conditions of their families, communities, and regions. *Processes* are the ways schools use their resources as expressed in curricular and instructional activities. *Outputs* or outcomes are all that the school tries to achieve and include the cognitive achievements of students as well as non-cognitive achievements such as the positive and negative feelings and attitudes that students develop relating to their activities, interests, and values.

Choice of model is important since a model attempts to identify the factors in schools that affect student learning. We thus have to ask: What if the factors identified as inputs and processes in functionalist-oriented input-output models are not the only ones of influence, or they are inadequately specified? Alternative models suggest that this may indeed be the case. Such models, for example, point to factors in the education process that are not adequately represented in input-output models and that might not only be regarded as important in their own right, but also for which there is empirical evidence that they affect student achievement. Among such factors are: the communitarian aspect of schools, which focuses on the school as a "small society," an organization in which informal and enduring relationships are driven by a common ethos; school policies and practices through which students are exposed to subject matter; student engagement, which involves the participation, connection, attachment, and integration of students into the school setting and its educative tasks, and for which personal relationships between students and teachers can act as a catalyst; and parent involvement expressed in assisting children's school learning or forming a functional community around the school (Lee, Bryk, and Smith, 1993).

Any model of schooling we may select will be limited in its ability to explain student learning. However, it is important to be aware of the nature of these limitations and to realize that they will be reflected in indicator systems that are linked to the model, as well as in any ameliorative action that might be taken on the basis of the information contained in the indicators.

USES OF NATIONAL ASSESSMENT DATA

The collection of data to construct and interpret indicators that represent the outcomes of the education system is expensive. It is thus reasonable to ask: To what use might indicators derived from a national assessment be put? There are various proposals in the literature on this topic (Greaney and Kellaghan, 1996). The following are eight such uses, which by no means exhaust the targets for national assessment that have been set in some countries.

Informing policy

The use of assessment data to inform public policy is very general and subsumes some other uses. It is argued that the kind of information available from national assessments provides an objective and hence sounder basis for policy decisions than other factors that contribute to policy making such as the personal biases of ministers of education or senior civil servants, vested interests of school managers or teacher unions, or anecdotal evidence offered by business interests, journalists, or politicians. Be that as it may, it would be foolish to imagine that policy formation can proceed without being influenced by these factors even when national assessment data are available.

We know relatively little about how the evidence obtained from national assessments influences policy in countries where such assessments exist, but it seems likely that it is used in much the same way as research evidence (Weiss, 1979). In some cases, the information may be directly acted on, as, for example, when a policy decision is made to increase the time allocated to a curriculum area or to introduce a new topic to a national curriculum. Examples of such decisions as responses to the findings of international assessments have been reported (Kellaghan, 1996a).

However, it is probably likely that national assessment information (as is the case with research) more frequently influences policy by entering the decision making arena as part of an interactive search for knowledge. In this case, indicator data are only one source of information and experience in the complicated process of policy formation. Alternatively, national assessment information may not be directly related to decisions at all but may serve to "enlighten" the policy making process, providing ideas about the extent and causes of problems and notions about appropriate solutions.

Monitoring standards

The monitoring of standards over time is frequently put forward as an important use of national assessment data. In light of this, many countries collect data on a cyclic basis—anything varying from every year to every ten years. The idea of monitoring standards is extremely appealing, especially when so many critics speak, invariably with inadequate evidence, about a decline in student achievement. However, national assessment data unfortunately are limited in what they can tell us about trends in achievement that might answer the critics.

First, the kind of achievement measured in national assessments may not be the kind of achievement that interests the business person or employer who is criticizing school graduates. Increasingly, critics are talking about the mismatch between the knowledge and skills produced in schools and the kinds of cognitive and noncognitive skills and competencies required for social and economic success in the contemporary world: higher-order thinking skills, analysis, problem solving, critical thinking, adaptability, team work, and flexibility. Criticisms will probably persist until it can be demonstrated that national assessments can provide evidence of such outcomes.

A second problem with the aspiration to use national assessment data to measure trends over time is that it is extremely difficult, some would say impossible, for technical and other reasons to do so (Goldstein, 1996). Certainly, changes in curriculum, in measuring instruments, and in the composition of student bodies all make it very difficult to compare student performances assessed at different points in time.

Given these difficulties, it is not surprising to find differences in the interpretation of national assessment data. For example, in speaking of the Chile national assessment, Olivares (1996) says that a comparison of 1988 and 1992 results points to an improvement in Spanish and mathematics. Himmel (1996), on the other hand, says that since no steps were taken to ensure comparability of tests from year to year, a comparison of changes in school performance over time is not possible.

Allocating resources

Some national assessments have been designed to help education managers make decisions about the allocation of resources. Thus, the results of the national assessment in Chile were used to identify schools with low achievement scores, and resources were then made available to these schools. Of eight thousand basic education schools in the country, about nine hundred from rural and poor urban areas were identified for intervention (Olivares, 1996).

While the use of national assessment data to allocate resources is attractive in principle, at least two disadvantages are associated with it. First, it necessitates assessment in all schools, not just a sample. And secondly, there is a danger that teachers will manipulate data collected in the assessment (e.g., in reporting the number of children from disadvantaged backgrounds) to improve their chances of obtaining additional resources.

Introducing realistic standards

The results of a national assessment, and even more so the results of an international assessment, can come as quite a shock to the educational community within a country. This happened in South Africa recently when the results of students' performance in mathematics and science in the Third International Mathematics and Science Survey (TIMSS), organized by the International Association for the Evaluation of Educational Achievement (IEA), became known. The performances of both black and white students were judged to be very poor by international standards.

There is a danger that information on student achievement (when compared to student achievement in other countries but not exclusively so) will be disheartening for countries that are less-developed economically and in which the resources available for education are much more limited than in wealthier higher-achieving countries. Obtaining information on student outcomes does not serve its purpose if it leads to fatalism or pessimism. One must weigh this danger against the value of having a realistic basis for addressing problems in the education system.

Participation in an international assessment has the advantage that it exposes participants to a wide range of experience and expertise that most likely will not be available within an individual country. It may thus contribute significantly to the building of capacity that may at a later stage be used in the design and execution of national assessments. Before deciding to participate in an international assessment, however, consideration should be given to a number of issues. First, in light of the relationship between the socioeconomic development of a country and the level of development of its education system, will the other participating countries provide a reasonable basis for comparison? Secondly, are the other demands on the education system (for example, the provision of primary schools in areas where there is an insufficient number) so obvious that expenditure on an international assessment would be very difficult to justify? And thirdly, if it is decided to participate in an international assessment, is it clear that resources will be available to address inadequacies in the education system that may be revealed in the assessment? If nothing can be done about the inadequacies, there would seem to be little point in spending a lot of money to obtain the information.

Identifying correlates of achievement

The value of descriptive statistics in themselves is extremely limited. Their value may, however, be considerably enhanced if they are placed in context. Indeed, some contextualization would seem to be a minimum requirement for interpretation. With this in mind, many national assessments collect data that not only put achievement data in context, but also may point to possible explanations of variation in achievement.

There are many examples of national assessments that identify correlates of achievement. For example, in a Colombian national assessment, a range of factors was found to be associated with achievement, including the emphasis given by teachers to specific curriculum areas, teachers' own educational backgrounds, students' living conditions, textbook-student ratios, and students' gender (girls performed better in Spanish; boys performed better in mathematics) (Rojas, 1996).

Two considerations seem relevant in the context of the identification of the correlates of achievement. First, there is a danger that an attempt will be made to collect too much data and that the boundary into the realms of what is traditionally considered to be research will be crossed. Second, there is a danger that correlations will be interpreted, without adequate supporting evidence, as providing evidence of cause and effect. Furthermore, interpretations relating to cause and effect may lead to decisions about manipulation. However, the existence of correlations between contextual variables and student achievement cannot provide unambiguous evidence of causal relationships, nor can it give any assurance on what the effects of manipulation of the variables might be. This is not to say that on the basis of other relevant evidence, considered in conjunction with indicator data, one may not be justified in proposing some kind of action. However, action should be taken with caution, and preferably on a limited pilot basis.

Directing teachers' efforts and raising students' achievements

On the assumption that schools would regard the content of a national assessment as reflecting important learning outcomes, Thailand introduced an assessment system that included outcomes not normally emphasized in Thai schools (Greaney and Rojas, 1996). Thai authorities may well have been correct in their expectation that the curriculum content of national assessment tests, and the information on student performance that those tests provide, would in time alter teachers' behavior. Furthermore, if teachers begin to emphasize and devote time to new curricular areas, we would also expect an effect on student achievement.

This line of reasoning is based on a long recognized fact that if assessment is regarded as important, it is likely to bring teaching and learning into line with what is assessed. One situation in which such information will be regarded as important is when the assessment can be regarded as high stakes: that is, when sanctions, either on teachers or on students, are attached to performance.

The testing of all students at particular grade levels in all schools in the assessment of the national curriculum in England and Wales is done partly with a view to directing teachers' efforts and partly with a view to improving school accountability. By contrast, national assessment in the United States was designed as an unobtrusive measure of the education system, limiting itself to describing what students know and can do without trying to influence directly what goes on in schools, though that view may be changing.

The implications of choosing one of the two models—one involving high stakes by linking sanctions to performance, the other not—should be given serious consideration when designing a national assessment. On the one hand, attaching high stakes to an assessment increases the likelihood that the assessment will affect school practice; on the other, high stakes can lead to a number of problems. First, the area that is assessed will be regarded as an important indicator of what is valued in education, leading perhaps to the neglect of other important curriculum areas. Secondly, schools may adopt strategies to optimize school performance on the assessment by using a variety of strategies, including refusing entry to low-achieving students, encouraging such students to leave the school, preventing them from taking the assessment by grade retention, or encouraging absence on the day of the assessment. Thirdly, schools will put considerable effort into test preparation. This can include training in test skills, choosing objectives based on items on the test and teaching accordingly, and presenting students with items similar to those on the test. As the match between instructional processes and test items increases, student performance on the test will indeed improve, but it will not be possible to say that the improved test scores are indicative of increased knowledge and skills. Indeed, empirical evidence indicates that improved performance on measures for which students are intensively coached is not matched by improved performance on other measures of achievement (Madaus and Kellaghan, 1992).

Promoting accountability

The use of high stakes, as discussed in the last section, raises issues of accountability. The use of national assessment data for accountability purposes raises other complex issues. Who is regarded as accountable will to some extent be a function of the level at which performance is reported. Thus in the British system of assessment, where performance was reported at the level of individual schools, schools can be held accountable. If individual teachers or schools cannot be identified, however, or if a sample rather than a whole population is involved in an assessment, poorly performing teachers or schools will not be identifiable and so cannot be held accountable.

When data are presented only at the level of the education system, it may seem reasonable to assign accountability to a government or ministry. It is for this reason that ministries may be reluctant to engage in a national assessment or, having carried one out, to release its findings.

Olivares (1996) has outlined some of the negative effects of a national assessment, National Educational Quality Assessment System (SIMCE), that were associated with accountability in Chile:

> There are teachers who develop their teaching programmes as if they are "preparing for the SIMCE" and there are even cases in which they have tried to "help" their pupils in unorthodox ways. More serious still is the fact that there are establishments where teachers have been asked to leave when the results do not match up to optimistic hopes. In the same way, there are heads of establishments who do not pass on their poor results to governors in case the latter could create difficulties. Finally there are the different sections of the social and school communities who tend to make a "league table of good and bad schools." (p.133)

Increasing public awareness

Although it may sometimes be expedient in the short term for a ministry of education to place limits on the amount of information it will make available to the general public, the long-term advantages of an open information system are likely to outweigh any short-term disadvantage. For one thing, the raising of public consciousness about educational matters is likely to increase public support for educational reform and for the funds that such reform may require.

While all eight of these uses have been put forward in the literature in favor of national assessments, it is difficult to know what uses in practice dominate the minds of policy makers or entice them to embark on an assessment. Certainly, several of the reasons given by Nwana (1996) in considering involvement in national assessment among African countries lack clarity. Uganda's ambition to monitor change over time and Burkina Faso's ambition to identify problem areas in the curriculum are clear enough.

However, the ambitions of Cape Verde to evaluate the education system's efficiency, of Lesotho to measure the "quality" of instruction, of Madagascar to improve the "quality" of education, and of Mali and Mozambique to assess the effects of innovations to enhance "quality" would require further specification before planning could commence.

SETTING THE SCENE FOR A NATIONAL ASSESSMENT

Once a decision is made to embark on a national assessment, it is important to do all that is possible to ensure that it is adequately supported, runs smoothly, and that its findings are seriously considered in making policy decisions. This will require that stakeholders are involved at an early stage and, insofar as is possible, that consensus on the assessment is achieved. One way of doing this is to establish a steering committee or independent governing board, representing various educational and community interests, which, though not essential for a national assessment, is strongly recommended by some commentators (Greaney and Rojas, 1996; Lapointe, 1990). The functions of such a committee would be to link the larger social and political aims of the national assessment with the more technical aspects of its implementation (Ilon, 1996). The committee would also provide status for the assessment; it would help ensure that the needs of the powerful national groups in the educational establishment are addressed; and it could help remove the administrative and financial stumbling blocks that can jeopardize or paralyze an assessment effort. The committee would provide overall direction as well as promote public awareness and discussion of results, thereby maximizing the impact of the assessment on educational policy making.

Because the educational-political power structures of countries differ, the interests represented in a national steering committee will vary from country to country. Obviously, representation should be provided for those responsible for administering the national assessment, those responsible for funding the exercise, those who will consider the results for policy making, and those who will be entrusted with the reforms that may arise from the assessment, such as school administrators and teachers. In general, the more homogenous a country and its education system, the easier it will be to establish a national assessment. Thus, establishing an assessment system should be less difficult in a country in which there is a single education authority and a uniform system of education than in a country with many education authorities (e.g., in a federal system) and in which there is not a national curriculum (Nwama, 1996).

Among the important issues that the steering committee should address are identifying the purpose and rationale of the national assessment, deciding on the content and on the grade levels to be targeted, developing a budget and assigning budgetary control, selecting an agency or agencies to conduct the assessment, determining terms of reference, and deciding on reporting procedures and publication.

Greaney and Rojas (1996) say that national assessments in Chile, Colombia, and Thailand benefited from commitment from ministries of education. In Colombia, the establishment of an assessment system was also probably facilitated by the enactment of a law in 1986 that established a division within the Ministry of Education with respon-

sibility for evaluating the quality of education in schools (Rojas, 1996). However, government commitment cannot be taken as a guarantee of positive effects. In England and Wales, the government was the driving force behind the assessment of the national curriculum. Nevertheless, the assessment evoked strong opposition from teachers who were concerned about the use of results to hold schools and teachers accountable for student performance. In one year, teachers actually refused to cooperate in the administration of the assessment, though the reasons for this were complex. Problems also arose in the national assessment in Egypt where local governments saw it as a threat to their autonomy and delayed its implementation.

CHOOSING AN AGENCY TO CARRY OUT THE NATIONAL ASSESSMENT

The execution of a national assessment requires expertise, or access to such expertise, in a range of activities including project management, curriculum analysis, test and questionnaire development, sampling, printing, distribution, data collection, processing and analysis, and reporting. Many ministers of education may look no further than their own personnel for such expertise. This was the case in Thailand, where the Office of Educational Assessment and Training, a section of the Ministry of Education, was given responsibility for the national assessment. In many other developing countries, some of the most knowledgeable educators may be employed within the ministry. Other reasons for basing an assessment in a ministry are that ministry personnel are likely to have ready access to up-to-date information for sampling purposes and that school inspectors or members of curriculum or textbook units should have considerable insight into key aspects of the education system. If a structure exists within the ministry for carrying out activities involved in a national assessment, this should help ensure that each phase of the assessment is adequately addressed. In Colombia, for example, there are separate units for implementing an assessment (including sampling, data collection, and data analysis), for research activities (including design of instruments and further analysis), and for dissemination (i.e., communicating results) (Rojas, 1996).

However, there are arguments against a ministry of education carrying out a national assessment on its own. Many ministries lack the required technical competence. Further, when a ministry carries out a national assessment, it may be slow to share information with others. Since ministry staff are likely to have a vested interest in the outcomes of an assessment, they might not be enthusiastic about focusing on potentially awkward issues or about making unpalatable findings public. For example, results that point to poor delivery of an education service or to failure by the formal education system to achieve a particularly sensitive goal (such as equality of achievement for ethnic groups) can embarrass ministry officials and (even more critically) their political masters.

A strong case, therefore, can be made for involving an external agency in the conduct of a national assessment. The case is supported by the fact that the main stakeholders in education may consider the information provided by a respected nongovernmental or independent agency more objective and thus more acceptable. Added to this is the fact

that technical competence is more likely to be found within university departments and independent research institutes than within ministries of education, though university departments do not normally have the capability of carrying out field work on the scale required by a national assessment. It was for reasons such as these that the national assessment in Chile was assigned initially to a nongovernmental body. In England and Wales, external agencies play the major role in the development of instruments, while schools themselves are responsible for local administration.

The use of an external agency, however, is not without its problems. In anticipation of these problems, a memorandum of agreement should be drawn up between the steering committee and the implementing agency before work begins. The memorandum should deal with such issues as funding, timetables, relations between the two bodies, and permitted data use.

An alternative to sole reliance on an internal or an external agency is to entrust the national assessment to a team composed of both ministry of education personnel and outside technical and curriculum experts. Such an arrangement can capitalize on the strengths of both groups and may increase the likelihood of general acceptance of the assessment findings. Various alliances are possible. In Mauritius, developmental work for the national assessment was undertaken by a semiautonomous Mauritian Examination Syndicate in collaboration with other national agencies, including the Ministry of Education, Science, and Technology; the Institute of Education; and its Curriculum Development Center. In Colombia, the Ministry of Education worked in cooperation with a number of public and private agencies. The design of instruments was contracted to the National Testing Service and data collection and analysis to the SER Research Institute (Rojas, 1996). In Chile, when national assessments were reinstated in 1988, the Ministry of Education contracted a university to carry out the assessment, but ministry staff worked with university staff over a period of three years, developing skills that were then available for later use. After that, university staff provided only advisory services (Himmel, 1996).

When necessary professional competence is not available locally, foreign experts may have to be employed. There are many examples of this. Data analysis for the Namibian national assessment was directed by Harvard University, while Florida State University provided assistance with aspects of sampling. When foreign experts are used, they should answer to the steering committee and should assist in the development of local capacity to conduct future assessments.

WHO WILL BE ASSESSED?

Inferences about the outcomes of an education system are based on an assessment of the achievements of students. To make such inferences, however, it is not necessary or desirable to involve students of all ages and grade levels in the assessment. Two decisions are required: one relates to the level of schooling that will be targeted; and the other to a choice between targeting students at a particular grade or age.

Choice of level of schooling

Policy makers want information on the knowledge and skills of students at selected points in their educational careers. Practically all countries that carry out national assessments target students in the primary grades. For example, Chilean students are assessed at grade 4 because this is the final grade of a subcycle of general basic education (Himmel, 1996). In most countries, national assessments are also conducted at some point at the secondary school level, usually at the lower or junior-cycle grades when education is still compulsory. To take Chile as an example again, students at secondary level are assessed at grade 8, which is the last year of compulsory education. Information at both levels can be valuable. Assessments at the primary-school level can identify deficiencies at an early point in the education system that indicate a need for remedial action. Information gained toward the end of compulsory schooling, or at a point when a large proportion of young people is still attending school, can also be useful if it provides some indication of how well students are prepared for life after school. In many developing countries, this will be at the primary school level.

Population defined by age or grade

In some national assessments (e.g., Chile and Scotland), only grade level is taken into account in defining the population for a national assessment. Many national and international assessments, however, use both student age and grade in their definition. For example, in the IEA literacy study, two populations were defined: students in the grade level containing the most 9-year-olds and students in the grade level containing the most 14-year-olds (Elley, 1992). In recent years in the United States, the grade level of the majority of students of a particular age has also been selected (Johnson, 1992). This strategy can be justified in industrial countries, where automatic promotion at the end of each grade is the norm, ensuring a pronounced link between grade and age.

In developing countries, especially in Latin America and in Francophone West Africa, the link between age and grade may not be close because of widely differing ages of entry to school and policies of nonpromotion. In this situation, students of similar age will not be concentrated in the same grade. To choose a population on the basis of age in such a school system would be disruptive since it would require students from several grade levels to take the tests at the same time. It would also be difficult to identify appropriate test content for the range of achievement that one would expect such students to exhibit. In the light of these considerations, a strong argument can be made for targeting grade level rather than age in national assessments in developing countries.

Total population or a sample of the population?

Having decided on the age or grade level that will provide information for a national assessment, the next issue to be addressed is whether all students at the identified age or grade will be assessed or whether only a sample will be selected for the assessment. Examples of both approaches exist. In Chile, Egypt, England and Wales, and France, all students are assessed. In most countries, however, a sample rather than the whole

population is assessed. Both approaches have advantages and disadvantages. The advantage of using the total population is that one has information on all students and schools which can be used for formative purposes (as in France, where results are used to help diagnose students' learning problems or in Chile, where results were used to identify schools in need of additional resources). The results can also be used for accountability purposes (as in England and Wales). Unless one requires information at the individual student or school level, however, the use of a sample may be more appropriate. There are, in fact, several advantages attached to this strategy, including reduced costs in gathering and analyzing data, greater speed in data analysis and reporting, and greater accuracy because of the possibility of providing more intense supervision of fieldwork and data preparation (Ross, 1987).

Whether the decision is to assess a total population or a sample, it is necessary to identify the precise population of interest. Since one is interested in the performance of students in the country at a particular age or grade level, it might seem appropriate to define the population in terms of students. However, this is unlikely to be practical. In most countries, a central agency such as the ministry of education will not have a list of all students attending school, which would be necessary to select a probability sample. And even if such a list existed, it would not be efficient or feasible to assess a sample of these students because, if randomly chosen, they would be spread over a large number of schools, making data collection difficult and expensive. Because of these conditions, schools (i.e., clusters of students) are usually identified as the population to be sampled in the first stage of sampling.

To select schools one needs an up-to-date list of all schools together with information that is relevant to stratification (e.g., size, location). In some countries, complete lists may not be available, or the lists may be too old to be useful. When lists are available, they should be checked carefully to see that all the schools actually exist. Even when a list of schools exists, one may have good reasons for not wanting to include all of them in the population for a national assessment. It may, for example, be decided to exclude from the assessment schools in which students are considered to be unassessable because of learning difficulties or of limited proficiency in the language in which the assessment will be conducted. Very small schools that could not on their own yield an adequate number of students may be clustered to form "pseudo schools," though this can be administratively complex and expensive. Because of the cost of data collection, schools in isolated areas may be excluded altogether. Exclusions should be kept to a minimum and information about them should be provided in the report of the national assessment.

Once the population of all schools eligible for selection has been identified, the next step, unless it is planned to assess pupils in all schools, is to select the schools in which students will be assessed. A variety of strategies is available for doing this, and the organization carrying out the national assessment may need external technical assistance in choosing the strategy and in deciding on the numbers of schools and students that are most appropriate. Great care has to be taken in this step because if the sample does not adequately represent the population, statements about national achievement levels of students will not be valid.

In sampling schools, it is common to stratify them according to such variables as location (area of the country, urban or rural); type (public or private); ethnic group membership; and religious affiliation. There are two reasons for this: stratification can improve the reliability of estimates; it also helps ensure that there are sufficient schools and students in the various categories, such as urban and rural, if differences between schools (and students) in these categories are of interest. To achieve sufficient numbers of schools and students, oversampling within some strata, rather than just selecting a sample size proportionate to the number of schools (or students) in the stratum, may be necessary. When strata are oversampled, a system of weighting will be required in aggregating student scores to ensure that the contribution of groups to aggregated statistics is proportionate to their size in the total population, not their size in the sample.

Following the selection of schools, the second stage of sampling requires that a decision be made about how students within a school are to be selected for assessment. Will all students in a school at the relevant age or grade level be assessed, or, if there is more than one class at the relevant level, will one class be randomly selected, or will some students be selected from all classes? Although the assessment of intact classes has obvious administrative advantages, the selection of students from several classes will provide a better estimate of the achievements of students in the school if the students have been assigned to classes according to different criteria or follow different curricula.

Determining the optimum size of a sample is not a simple matter. What one is seeking to do is to obtain the required level of precision of estimates within the resources that are available for data collection. As a rough guide, one may consider the sample size used in TIMSS. In each country, a sample of one hundred fifty schools was selected and within each school, thirty students were selected. Circumstances may require a larger sample, however. This is likely to be the case if a country is divided into a number of administrative regions, if one wants to be in a position to compare performance in a number of sectors of the education system, or if large differences in mean achievement between schools are anticipated.

It is not necessary that all students take the same test in a national assessment. Broad curriculum coverage may be attained by the use of booklets containing different sets of items. Each pupil will respond to only one booklet. With this system, class averages can still be estimated for each task and aggregated to provide summary statistics. This procedure is followed in the United States, where samples of students are administered one-seventh of the total number of test items developed for each grade. Such sampling, known as matrix sampling, permits the coverage of much larger sections of the curriculum and may prove less time-consuming than administering the same test to all students. However, the technical and logistical requirements, including printing many different forms of a test, packaging and administering them, and combining test results may be daunting, especially in a country's first national assessment.

The cooperation of schools (which in most countries can decide whether or not to participate in a national assessment) is necessary if the designed sample is to be

achieved, which, in turn, is necessary to make statements about the education system in general. Inevitably, one will be faced with the issue of how to deal with nonresponse, caused either by a school's refusal to cooperate or by student absence on the day of testing. There is no absolute figure for participation that one can regard as ensuring that the sample of schools and students that responds adequately represents the performance of the target population as a whole. Any level of nonresponse must be a matter of concern. However, if one is to be realistic, one will accept that for a variety of reasons, not all schools that are invited to participate will do so. As a general guide towards what might be regarded as an acceptable nonresponse rate, one might consider the standards set for TIMSS. In that international study, the minimum acceptable school-level response rate, before the use of replacement schools, was set at 85 percent. The same figure was set for the student response rate within schools (Martin and Kelly, 1996). Replacement schools may be used to ensure adequate sample size. They do not, however, increase response rate and cannot be taken as evidence that any possible bias that may have arisen from the decision of schools not to participate is removed. The report of a national assessment should provide information on response rates for schools and students, as well as the extent to which replacement schools were used.

When a complex sampling design involving such procedures as stratification, clustering, and weighting is used to select participants for a study, the effects of the design on sampling error have to be taken into account in analyses. Otherwise, the estimate of true sampling variability will be biased. Reducing this bias is not a simple procedure. One approach with large-scale survey data when formulae are not readily available for the calculation of sampling errors is to use a "jacknife" procedure. The procedure requires that estimates of statistics be made on the total sample of data, following which the data are divided into groups and calculations are carried out on reduced bodies of data in which subgroups are omitted in turn. The mean of the subsamples is then estimated and its variance is taken as an approximation of the subsample estimates. An alternative method of estimating error involving multi-level modelling is also available. It is regarded as statistically more efficient and also has the advantage that it provides information about variation among schools (Woodhouse and Goldstein, 1996).

WHAT IS ASSESSED?

Both political and technical considerations affect the identification of the knowledge and skills to be examined in a national assessment. The role of political factors is evident in the need to select content that addresses the informational requirements of key policy makers. Technical considerations are apparent in deciding what is technically possible to measure and in evaluating cost and logistical requirements. In practice, tensions that have to be dealt with are likely to emerge between the informational needs, goals, and ambitions of those commissioning a national assessment and the ability of the assessment to accommodate them, given financial, technical, administrative, and time constraints.

Information normally collected in national assessments can be divided into three main categories. First, all assessments measure *cognitive outcomes of instruction*—specifically,

competence in areas of the curriculum. Secondly, in recognition of the view that education should not be confined to the development of cognitive abilities, many national and international assessments collect data on *noncognitive outcomes*, including self-concept, attitudes, and values. Thirdly, most national assessments also collect contextual information on *background variables* such as school and nonschool factors that may contribute to student achievement.

Cognitive outcomes

All countries that conduct national assessments examine the students' first language and mathematics. Science is sometimes included and, in a smaller number of countries, a second language, art, music, and social studies (Kellaghan and Grisay, 1995). The attention to language and mathematics is an indication of the importance of these subjects for basic education and merits serious consideration by any country embarking on a national assessment, since the primary concern of such an assessment should be to collect data that provide information on the extent to which important goals of the curriculum are being achieved.

Once a particular curriculum area has been chosen for assessment, it is necessary to ensure that the techniques used to obtain information on student achievement are comprehensive in their coverage of curriculum. Comprehensive coverage ensures that the results of a national assessment provide an accurate picture of student performance and identify the particular strengths and weaknesses of the curriculum as reflected in student achievement profiles.

Not only should the whole curriculum be covered, its various domains should be sampled in sufficient detail to allow inferences to be made about the extent to which each area is being taught and learned in schools. For example, the TIMSS assessed six areas of the mathematics curriculum in elementary schools: whole numbers; fractions and proportionality; measurement, estimation, and number sense; data presentation, analysis, and probability; geometry; and patterns, relations, and functions (Mullis, Martin, Beaton, Gonzalez, Kelly, and Smith, 1997) and four areas in science: earth science, life science, physical science, and environmental issues and the nature of science (Martin, Mullis, Beaton, Gonzalez, Smith, and Kelly, 1997). The differential information that can be obtained by ensuring that each content area is adequately sampled provides a good basis for determining areas of the curriculum that are adequately implemented and achieved and those that are not.

Noncognitive outcomes

Students' development is multifaceted and schools, in addition to promoting development in cognitive areas, will also seek to enhance development in the personal and social skills of students, such as those involved in self-concept, cooperation, leadership, innovation, self-confidence, motivation, and independence.

It was with considerations such as these in mind that some national assessments have included measures of noncognitive outcomes. The Colombian national assessment included measures of student attitudes toward school, subjects, and teachers; creativity; self-esteem; and democratic values, in addition to measures of Spanish and mathematics (Rojas, 1996). In Chile, questionnaires were administered to assess student self-concept, attitude toward school and learning, peer and social relationships, vocational orientation, and value acquisition (Himmel, 1996). At the international level, the IEA reading literacy study evaluated student attitudes toward reading by examining the extent of voluntary reading of books, comics, and newspapers; book reading preferences; and the amount of encouragement students received to read and to use the library (Elley, 1992). Although noncognitive outcomes would appear to be an important aspect of education, their measurement and interpretation have proved problematic and, in the case of Chile, their use was judged unsuccessful (Himmel, 1996).

Background variables

Reference has already been made to the need to have a model of the education system in mind when conducting a national assessment. An important part of such a model is likely to be information on background variables—both in and outside the school—that will be used in the interpretation of data on the performance of students on cognitive or noncognitive achievement variables.

In an attempt to obtain information that would promote understanding of the factors that determine the achievement of minimum learning objectives, the Performance Evaluation Program (PER) that was carried out in Chile between 1981 and 1984 obtained data on five sets of background variables relating to student background; teachers (training, experience) and classroom variables; school principals and schools (including expectations for learning, administration, and discipline); local environments (including socioeconomic level, educational resources); and institutions (structure of educational system, educational policies, financing) (Himmel, 1996).

In Thailand also, an attempt was made to identify factors related to the scholastic achievements of students, such as socioeconomic status, school size, grade repetition, and access to preschool (Chinapah, 1992). In Colombia, data were collected on student gender, educational history (including grade repetition), length of school day, text book availability, time devoted to homework and watching television, teacher characteristics such as level of formal education and the amount of in-service training received, and home background factors (e.g., facilities at home, family size, education and work expectations, and quality of dwelling) (Rojas, 1996).

THE ASSESSMENT INSTRUMENTS

Achievement variables

A number of general points may be made about the development of assessment instruments, a highly technical task that is likely to be assigned to the agency responsible for

implementing the assessment. First, the content of instruments must be consistent with the overall objectives of the assessment. If one objective, for example, is to measure competence in the mathematical domains of computation, concepts, and problem-solving skills, the assessment instrument design must ensure that each of the three domains is adequately assessed. Second, policy considerations should always be kept in mind. This implies that coverage of the curriculum in the instruments must allow inferences to be made about the extent to which curriculum objectives are being achieved in schools.

Documents that specify a curriculum or syllabus, when available, will be an important source in developing instruments. Such documents, however, may not provide sufficient detail or indication of the relative importance of curriculum content, in which case recourse will have to be made to other sources, such as textbooks and teacher experience.

Experience indicates that a table of specifications can greatly facilitate the development of multiple-choice or short-answer assessment instruments (Bloom, Madaus, and Hastings, 1981). A typical table consists of a horizontal axis that lists the content areas to be assessed (for example, aspects of the mathematics curriculum for a given grade level such as whole numbers, fractions, measurement) and a vertical axis that presents in a hierarchical arrangement the intellectual skills or behavior expected of students (e.g., computation, understanding). Cells are formed at the intersections of the two axes. It is the responsibility of test developers to assign test items or questions to each cell based on their perceptions of the relative importance of the objective represented by the cell. Cells are left empty if the objective is considered inappropriate for a particular content area. Table 1 provides an example of a table of specifications developed for a mathematics curriculum for the middle grades of primary school in Ireland.

If there are plans to repeat a national assessment at a later date to monitor trends in achievement over time, the same test (or a portion of it) will have to be used again. In this situation, the assessment instrument should not be made public, and all copies should be collected immediately after test administration. Examples of items used in the assessment procedure may be made public so that school personnel know what is expected of students.

Type of test

Most national and international assessments rely to a considerable extent on the multiple-choice test format in their instruments. A multiple-choice item usually consists of a statement, direction, or question followed by a series of alternative answers, one of which is correct. The advantages of such items include speed of response, ease of marking or correcting, objective scoring, potential for covering a considerable portion of a content area, and high reliability or consistency (Frith and Macintosh, 1984). If optical scanning machines are available, multiple-choice answer sheets can be scored and processed quickly, though in countries where labor is cheap, hand-scoring may remain a viable alternative. When multiple-choice data are computerized, it is easy to

provide detailed feedback on characteristics of individual items, objectives, and levels of achievement of students classified by, for example, grade, age, gender, and ethnic or linguistic affiliation, if information on these variables has been collected in the survey.

There are also negative aspects to multiple-choice tests. Their construction is time-consuming and expensive. They cannot be used to assess important aspects of the curriculum, such as oral fluency, writing, and practical skills. They have also been criticized for overemphasizing the factual at the expense of determining the student's understanding of the content being assessed.

Table 1. Table of Specifications for Mathematics Test (Middle Primary Grades)

	Content area						
Intellectual behavior	Whole numbers	Fractions	Decimals	Measurement	Geometry	Charts and graphs	Overall total
Computation							
Knowledge of terms and facts							
Ability to carry out operations							
Total							
Concepts							
Understanding of math concepts							
Understanding of math principles							
Understanding of math structure							
Ability to translate elements from one form to another							
Ability to read and interpret graphs and diagrams							
Total							
Problem solving							
Ability to solve routine problems							
Ability to analyze and make comparisons							
Ability to solve nonroutine problems							
Total							

The many disadvantages of multiple-choice tests, particularly in the context of their use in evaluating the output of schools (Madaus and Kellaghan, 1992), point to the need to supplement their use with other forms of assessment. Thus, for example, in addition to items that simply require a student to identify a correct answer among a number of possible answers, the student may be provided with no options and so have to construct the answer. Other item types that require the student to construct a response include those in which the student has to write a word, phrase, sentence, or even an extended essay in which he or she has to organize and present thoughts in a coherent, and perhaps persuasive, fashion. Some use has also been made of more complex practical assessment, such as requiring a student to set up an experiment in science.

In recent years, there has been much talk of the need for what is called "performance" assessment which involves tasks requiring the student to demonstrate knowledge and skills and construct responses to complex tasks. Such tasks are used to assess competency in such areas as practical measurement skills in mathematics or in conducting a scientific experiment. Ideally, a performance assessment should provide information on the procedures that students use, their ability to use implements, and the quality of a completed product.

While performance tasks might have several advantages—for example, in providing clear models of acceptable outcomes, positively influencing learning and instruction, and encouraging the teaching and learning of higher forms of mental functioning—they present serious problems because of the possibility of variation in their administration and scoring procedures (see Gipps and Murphy, 1994; Mehrens, 1992; Meisels, Dorfman, and Steele, 1995). The most comprehensive approach to the use of performance testing in national assessment is to be found in the British national assessment system. The keys to the system, as it was originally envisaged, were Standard Attainment Tasks (SATs), which were designed to provide information on students' performance on a cluster of attainment targets that had been set for a range of curriculum areas. The tasks used a wide range of modes of presentation (e.g., oral, written, pictorial, video), operation (e.g., mental, written, practical, oral), and response (e.g., multiple-choice, writing a short prescribed response, open-ended writing, a practical product). Teachers were required to integrate the tasks into their normal classroom practice, thus avoiding the artificial separation of assessment and teaching. They also scored the students' performance.

Experience with the first major assessment of 7-year-old children in 1991 brought to light serious inconsistencies in the administration and scoring of the SATs (Madaus and Kellaghan, 1992). The lack of standardization that was a feature of administration and scoring must call into question the use of the data obtained for comparing individual pupil scores or aggregated school scores. As a result of the problems experienced in administration of SATs relating to topic effect, rater effect, and the generalisability of scores, assessment procedures have been greatly modified and will in the future involve more streamlined and conventional tests (Kellaghan, 1996b).

Nonachievement variables

Questionnaires and rating schedules designed to provide contextual and policy-relevant information can be administered at the same time as the achievement instruments at relatively little additional expense. Contextual information might include information about teachers (e.g., their qualifications and frequency of attendance at courses); class size; length of school day; teaching time; school facilities (e.g., number and condition of desks and books); the amount of the textbook covered during the school year; time devoted to curriculum areas; amount of homework assigned; percentage of students being tutored outside school; and the attendance, completion, and promotion rates of students.

Identification of contextual factors related to student achievement can be particularly useful for policy makers, who can use this information to influence the reallocation of scarce financial resources. Knowledge of contextual variables can forestall policy makers' tendencies to focus on a single variable without considering other possible factors that might account for a finding. It can also help in the identification of manipulable variables—for example, the time allocated to curriculum areas, the nature of preservice and inservice teacher training, and student promotion rates—that appear to be positively related to student achievement.

ADMINISTRATION OF THE NATIONAL ASSESSMENT

The logistics of administering a national assessment are complex. Targeted schools have to be contacted to secure their cooperation; materials have to be printed, packaged, and distributed; personnel to administer the assessment have to be recruited and trained; supervisory visits to assessment centers have to be organized; and answer sheets and questionnaires have to be retrieved from schools, cleaned, scored, and matched. Given the amount of data a national assessment will generate, it is useful at an early stage to set up a database to which data can be added when materials are returned from schools.

There are two possible approaches to the administration of assessment tasks in schools. One is to have teachers administer the tasks in their own schools; the other is to use outside staff to visit schools.

Entrusting as much as possible of the actual test administration to teachers in the schools in which the assessments are being conducted will reduce administrative costs substantially. Teacher involvement may also contribute to the assessment's political viability and increase the probability that reforms prompted by the assessment will be acted on. However, there is also a down side to the use of teachers. All teachers may not follow administration procedures adequately, giving rise to problems in comparability.

Where there is a serious concern that the validity of an assessment may be compromised by assigning test administration to teachers, alternative strategies should be adopted.

Use of the ministry's inspectorate and curriculum staff or researchers to administer tests and other instruments may be a viable alternative, and one more likely to ensure that standard procedures are followed in all classrooms. Furthermore, the involvement of ministry personnel confers a certain status on the exercise. However, there are also disadvantages associated with this procedure. First, there is the question of the cost of the invigilators' time and perhaps of travel and subsistence. Secondly, it may disturb teachers and students. And thirdly, there is the question of time scale. Since assessments should be carried out in all schools at about the same time, a large number of external assessors would be required to cover the whole of the country in a limited period of time.

Whoever is responsible for administering assessment procedures, it is important that the conditions under which the assessment is conducted are as uniform as possible from school to school. As well as ensuring that all students are administered the same or equivalent tasks, there should also be uniformity in instructions and materials, in the time allowed to complete tasks, and in the general assessment environment.

There are a number of ways of to deal with the problem of variation in procedure. One is to make assessment tasks unambiguous and relatively simple. Secondly, training should be provided in the administration of the tasks. Thirdly, a detailed manual for the use of those who will be responsible for administration should be provided. The manual should be clear about the number of students to be tested and the method of selecting them if they have not been preselected by the implementing agency. Instructions should contain a work schedule and precise details for administering instruments and tasks. A sufficient supply of materials should be available. Finally, some kind of quality control of the administration of an assessment is also desirable. This would involve a number of individuals who are thoroughly familiar with the required procedures visiting a sample of schools during the administration of the assessment to ensure that procedures are being followed. Obviously, such a task has to be carried out with great sensitivity and, even when it is, teachers may resent it.

FREQUENCY AND TIME OF ASSESSMENT

The frequency with which a national assessment is carried out varies from country to country. In France, all students in grades 3, 6, and 10 are assessed every year and a sample of grade 9 students is assessed about every five years. In England and Wales, all students (at 7, 11, and 14 years of age) are assessed every year. Elsewhere, assessments are less frequent. For example, Canada follows a three-year cycle, Finland a ten-year cycle. In some countries, there is no predetermined cycle and assessments are carried out when considered necessary.

Unless feedback is provided to individual schools (as in England and Wales) or for individual students (as in France), an annual assessment would not seem to be necessary. If the purpose of the assessment is to provide information on the performance of the system as a whole, an assessment in a particular curriculum area every three or five years would seem adequate. Educational systems do not change rapidly and more frequent assessments would be unlikely to register change.

Although assessment in a curriculum area may be carried out only every five years, this does not mean that assessments in other curriculum areas might not be carried out in the intervening years. A national assessment system in which, for example, basic curriculum areas (reading, mathematics, science) are assessed every three years might involve the assessment of all three areas together every three years or it might involve assessing one curriculum area every year.

The time of year an assessment is carried out will to some extent be determined by its purpose. If, for example, the purpose is to obtain information on student achievements in the last year in which most students share a common curriculum (as in the case in Canada when 13-year-old students are assessed) or in the last year of compulsory education (which is age 16 in most Canadian provinces), then an assessment will be conducted towards the end of the relevant school year.

If, on the other hand, the results of a national assessment are to be used for diagnostic purposes and are expected to directly affect teaching in individual classrooms, then students will be assessed at the beginning of the school year. This is the case in France, where students are assessed at entry to lower secondary school (grade 6) and at the beginning of upper secondary school (grade 10). A country may carry out assessments at different times of the year, depending on the kind of information that is required. Thus, in France, in addition to beginning-of-year assessments, an assessment at the end of the year in the final grade (grade 9) of lower secondary schooling is carried out to provide summative information on how well the system is performing at this point.

DATA ANALYSIS

When assessment tasks have been completed in schools, all materials relating to them should be returned to the implementing agency. A data management system should be in place to receive materials, checking that nothing is missing. It may be necessary to contact schools that have not returned all their materials to determine the reasons for the delay and perhaps to encourage them to carry out the assessment if they have not already done so.

When materials are all returned, student responses will be scored and recorded. Data will be cleaned and entered into a database, which might have been established at the beginning of the study or, if not, will need to be established at this stage.

The form of analysis that is carried out will depend on what decisions have been made about how results should be reported (discussed in the following section). When data other than achievement data have been collected, analyses should be designed to identify relationships between student achievement and nonachievement variables, which will probably include personal characteristics of students and school facilities. The results of such analyses can help prevent people from arriving at simplistic conclusions—that, for example, private schools are "better" than public schools when differences in students' home backgrounds may contribute substantially to the difference.

Comparisons of assessment results of schools of different types, regions, or groups (e.g., ethnic groups) are likely to be of great interest to policy makers. Certainly, differences between boys and girls or between students in urban or rural schools should be a cause of concern to policy makers and education managers. Indeed, one of the purposes in carrying out a national assessment may be to obtain empirical data on the extent of such differences. However, such results should be submitted to considerable scrutiny and not be interpreted simplistically to imply causation when it is not warranted. Further, it would be naive to regard the mere recognition (and possible publication) of differences as an adequate response to the problems that may underline differential performance.

REPORTING THE FINDINGS

Assessment results should be reported as soon as possible after data collection. If they are delayed beyond seven or eight months, the usefulness of the exercise is diminished. Reports should be concise, simply written, and devoid of educational jargon. The timely, well-presented, and well-illustrated reports produced by the NAEP, and recently by TIMSS, can serve as models. One should consider the publication of summary reports for the general public and more detailed reports for policy makers, education managers, and teachers.

Many approaches have been used in national and international assessments in reporting results. One involves reporting average levels of student performance in a curriculum area. The others involve reporting the percentage of students associated with specified achievements. The achievements, however, are defined in different ways.

Average performance of students in a curriculum area

If the individual scores of a representative sample of students in a country are added and then divided by the number of students, one gets an overall average for performance in a particular curriculum area, at a particular age or grade level, for that country. The procedure may not be quite as simple as this in practice, since adjustments may have to be made to take account of disproportional sampling of students in different geographical regions or types of school. The basic point, however, is that one is seeking to represent in quantitative terms the average level of performance in the country.

This information is of limited value for a number of reasons. First, it does not tell us whether the average obtained can be regarded as "satisfactory" or "unsatisfactory" unless comparative data are available with which to compare the obtained average score. Thus, for example, the information could be used as a general indication of whether standards in the country were stable, rising, or falling, as long as comparable information were available from an earlier point in time. It would also be useful if similar information were available from other countries, as is the case in international studies of assessment. Both the IEA and the International Assessment of Educational Progress (IAEP) have reported mean scores for participating countries in a variety of curriculum areas. OECD has made use of these data to highlight differences in achievement among its member countries (OECD, 1995).

A second problem with using average aggregate scores for a broad curriculum area such as mathematics or science is that in adding up the number of correctly-answered items, one masks differences in performance that may exist between different domains of the subject, e.g., in mathematics, between computational ability and problem-solving ability. Thus, it is useful to report average scores for domains, in which items of the curriculum are grouped in a meaningful way, since differences between performances in domains can be of diagnostic value.

Mean scores for sectors of the education system can also be of interest. In the Colombian national assessment, for example, mean achievement scores were calculated for each curriculum area by state, location (urban/rural), and type of institution (public/private) (Rojas, 1996).

Percentage passing items

Some national (e.g., the U.S. NAEP) and international (e.g., the IAEP) assessments have reported results at the individual item level. For each individual item, the percentage of students answering correctly was reported. Average percent-correct statistics were then used to summarize the results (see Baker and Linn, 1995; Phillips et al., 1993). Such a form of reporting has value, though it probably is too detailed for most readers. Furthermore, if comparisons are to be made from one assessment to another or among the results for different grades, the approach requires that identical sets of items be used.

Percentage achieving mastery of curriculum objectives

In another approach, the percentages of students who achieve mastery of major curriculum objectives are presented. This approach is an intermediate step between reporting item statistics and statistics for broad domains within a curriculum area. In one Irish assessment, the mathematics curriculum for students in grades 5 and 6 was divided into 55 objectives in computation, concepts, and problem solving. Objectives called, for example, for the student to be able to add a column of numbers containing not more than five digits; subtract two numbers containing not more than five digits; perform simple arithmetic operations involving zero; and identify common factors between two numbers. A student was regarded as having mastered an objective when he or she correctly answered a specified number of items per objective on a multiple-choice written test. Statistics were provided for each of the 55 objectives, indicating the percentage of students who had mastered the objective. Aspects of the national curriculum that posed problems were identified (Kellaghan, Madaus, Airasian, and Fontes, 1976).

Percentage achieving specified attainment targets

In some education systems, specific attainment targets are set for students at varying points in their educational careers. Where this is the case, an assessment system may be designed to obtain estimates of the number of students who are reaching these targets. In the British system, the extent to which students are meeting attainment targets of the national curriculum at ages 7, 11, 14, and 16 is identified. Each target is divided into

levels of ascending difficulty on a scale of 1 to 10, with clear criteria defining what a student must know, understand, or be able to do to be rated as scoring at that level. There were thirty-two targets relevant to 7-year-olds in the 1991 assessment: five in English, thirteen in mathematics, and fourteen in science. Results were reported as the percentage of students who satisfied each level in each curriculum area.

In English, the percentages attaining levels 1, 2, and 3 were given for five targets: speaking and listening, reading, writing for meaning, spelling, and handwriting. In mathematics, examples of targets for which results were presented were number, algebra, and measures; using and applying mathematics; number and number notation; number operations (+, −, ÷, x); and shape and space (two- and three-dimensional shapes). The science targets included life processes, genetics and evolution, human influences on the earth, types and uses of materials, energy, and sound and music (Great Britain, Department of Education and Science, 1991).

Percentage functioning at specified levels of proficiency

Another way of presenting results, used in several state and national assessments, is to construct a proficiency scale through statistical procedures and to determine levels on the scale through judgmental processes. Proficiency scales have been constructed for both national (Canada, United States) and international (IEA, IAEP) assessments. For example, five proficiency levels were established in mathematics, science, reading, and writing in the Canadian national assessment of 13- and 16-year-olds (Canada, Council of Ministers of Education, 1996). Each level is described in terms of the knowledge and skills that a student operating at the level should exhibit; the percentage of students functioning at each level is then reported. In science the student should be able to describe, at a given level:

- Level 1: Physical properties of objects
- Level 2: Qualitative changes in the properties of a substance when heated or cooled
- Level 3: The structure of matter in terms of particles
- Level 4: Qualitatively, a chemical reaction or phase change
- Level 5: Quantitatively, the product of a reaction given the reactants, or vice versa

Sometimes labels are attached to levels. For example, students in the NAEP are described as lacking basic competency, as having attained basic competency, as being proficient, or as being advanced. An alternative nomenclature was used in an assessment in Kentucky: students were described as novice, apprentice, proficient, or distinguished (Guskey, 1994). Although such labels have obvious intuitive attractions, they can have negative connotations. They can also mean different things at different grade levels, and even for the same grade they are likely to be interpreted in different ways by different people. When results are reported as levels, it would seem preferable to avoid labels, using instead verbal descriptions of what a student at a level knows or can do.

There may be more serious problems associated with scaling. One is the assumption that student responses in an assessment are determined by a single "trait" value and that,

as a consequence, the set of items can be regarded as unidimensionally reflecting "reading ability," "mathematical ability," or "scientific ability." There are reasons, arising from a consideration of the nature of achievement and of the statistical procedures involved, that lead one to question the validity of this assumption (see Goldstein, 1996).

The practical effects of scaling procedures on curriculum representation and balance can be seen in attempts to set achievement levels (basic, proficient, and advanced) for the United States 1990 NAEP mathematics test at grades 4, 8, and 12. In considering these levels, the National Assessment Governing Body was struck by the inadequacy of the item pool and, in particular, by the lack of what they described as "sufficiently challenging" items. What had happened was that more difficult items that did not meet scaling criteria had been excluded from the item pool, since they did not contribute to the scale, thus reducing the congruence between the assessment procedure and the curriculum domain it was designed to represent (Kellaghan and Grisay, 1995). The question to be addressed in this context is: Does one accept changed curriculum coverage in the interest of meeting technical standards or does one maintain the position that coverage is paramount?

The distribution of achievement

In addition to information on mean achievement, it is useful to have information on the distribution of students' achievements in a curriculum area. Such information focuses on disparities between high and low achievers. The use of proficiency scales—if one accepts their validity—will do this, but even if such scales are not used, an examination of raw data, coupled with a judgmental process regarding the adequacy of students' performance, may provide useful insights. Disparities in achievement may be analyzed by gender, geographical location, or type of school to provide a richer reading of the data thrown up in the national assessment. It may be, for example, that one finds a higher incidence of low scorers in some areas of the country than in others and among boys than among girls.

ESTIMATING COSTS

It is not possible to estimate definitive costs of a national assessment that would apply to all situations, as these will vary depending on the level of socioeconomic development of a country and the characteristics of the assessment being carried out. One would expect higher costs for salaries and services in an industrialized country than in a developing country, though developing countries may have to import some of the expertise required, which can of course be costly.

The available data also support the view that estimates of the cost of national assessments vary considerably. In Jamaica, for example, it has been estimated that the cost per student assessed is US$1.00 for 20,000 test takers and US$1.32 for 10,000 test takers (Ilon, 1996). The estimated cost for Chile's Performance Evaluation Program (PER) (1981-84) was US$5.00 per student (Himmel, 1996). For the more recent (1988)

National Educational Quality Assessment System (SIMCE) the cost is estimated as US$4.31 per student at grade 4 and US$6.94 per student at grade 8 (Olivares, 1996).

Since the conditions and the circumstances in which an assessment is carried out will affect its cost, it is necessary to estimate the cost of each assessment before it is carried out. Ilon (1996) has outlined the considerations that need to be taken into account in this exercise, which is likely to involve assessment specialists, policy makers, educators, and economists. These individuals will have to collaborate in an interactive process in which the proposed components of the assessment are modified until costs fit within a budgeted amount. Those involved in the process will soon realize that each component in the assessment will involve decisions that weigh costs against the information that stakeholders would like to obtain. One may have to compromise on what is desired in the light of budgeting constraints. Ilon raises a range of questions relating to the various components of a national assessment that need to be addressed when considering costs.

Steering committee

The cost of a steering committee is likely to be small relative to total costs. Costs can, however, vary depending on the extent of transportation and administration required, and whether or not members are compensated for their time. Transport costs obviously depend on how geographically dispersed numbers are while the issue of compensation will depend on tradition.

Implementing agency

Costs will vary depending on whether the agency has the necessary facilities, experience, and expertise to carry out the national assessment, or whether it will have to upgrade its facilities, provide additional training for staff, and/or contract outside consultants, all of which can involve considerable expense.

Building support

It may be that support for a national assessment is relatively widespread in a country, in which case only a simple information pack may be required. However, if there is a need to raise consciousness about the value of a national assessment and to enlist the support of teachers, parents, and the general public, considerable expense may be incurred both in employing people to produce literature, videos, and posters and in conveying the message through meetings, talks, radio, or television.

Target population

Several factors relating to the choice of the target population have implications for cost. First, the number of grade or age levels to be assessed is relevant, not just in the administration of an assessment procedure, but also in the cost of the development of instruments. Secondly, costs will be affected, both in administration and in production of materials, if it is decided to test a whole population at a particular age or grade level

rather than a sample. Thirdly, selecting a population that strands different types of school will add considerably to the cost. For example, if the target population is 13-year-olds and children of this age are found in both primary and secondary schools, it will be necessary to sample an adequate number of both primary and secondary schools to obtain an accurate picture of the achievements of the relevant population. Fourthly, targeting an age level is likely to be more expensive than targeting a grade level since students of any particular age may be spread over a number of grades, which may in some circumstances require additional testing sessions. However, this cost is more likely to be borne by the school than by the agency implementing a national assessment, unless the latter is paying for invigilation.

Instrument content and construction

The various options for selection of the content and form of assessment should be considered in terms of cost as well as in terms of other considerations such as validity and ease of administration. There are cost implications in the construction of assessment procedures, their administration, and their scoring. Multiple-choice tests are relatively expensive to construct but relatively inexpensive to score. The same is true of performance assessments that involve the use of a tape or a video and to which the student is asked to record his or her response. In the case of such techniques, however, there may be costs in terms of equipment (e.g., a tape recorder) that schools may not have and the transportation of the equipment to schools. Other forms of performance assessment in which students are required to write an extended essay or to demonstrate some practical knowledge or skills in the presence of an assessor (e.g., to carry out an experiment in science, to play a musical instrument) may not be very expensive to design but are the most expensive to administer and score.

Administration manuals

Manuals used to ensure uniformity in the administration of an assessment can vary in their form of presentation (e.g., paper quality, use of color, and binding), but overall, the preparation of the manual should not be a major cost. The cost of distribution should not be forgotten.

Administration

Data collection will be the most expensive item in a national assessment. It involves a number of steps. Information may be obtained from schools in advance of the assessment about their eligible students from which the implementing agency will choose those required to participate in the assessment. Test materials have to be designed and printed and then sorted, packed, and sent to schools. Arrangements will also have to be made for the return of the materials and to cover the cost involved. A major factor to be considered in data gathering is whether special invigilators will be required to visit schools or whether teachers will administer the tests in their own schools. In either case, but particularly if teachers administer the tests, provision will have to be made for some supervision of the administration of the assessment tasks.

Analysis

Analytic costs will depend on the type of assessment procedures used and the availability of technology for scoring and analysis. While machine-scoring of multiple-choice items is normally considered to be much cheaper than hand-scoring, this may not be the case in a country where technology costs are high and labor costs are low. The cost of the analysis that will produce the information on student levels of achievement will depend on how it is planned to report results. The most expensive method involves scaling and reporting in terms of proficiency levels.

Reporting

Funds must be set aside for the writing of a report and the possibility must be considered that different versions of the report may be required for government, teachers, and the general public.

Cost components

As noted above, costs of a national assessment are likely to vary from country to country. Although the proportions of cost attaching to varying components may also vary, the proportions involved in the U.S. NAEP may provide a rough guide in the costing exercise. Data collection (30 percent of total cost) was the most expensive item, followed by instrument development, data analysis, and reporting and dissemination, each of which required 15 percent of funds. Sampling and selection (10 percent), data processing (10 percent), and governance (5 percent) were the least expensive elements (Koeffler, 1991).

Loxley's (1992) advice to set aside a contingency fund for emergencies, although intended for those involved in international assessment, is also relevant to national assessment. In recommending that 10 percent of the budget be earmarked for this purpose, he notes that "it is never a question of whether emergencies will arise, but rather of when and how many" (p. 293).

CONCLUSION

Few people would doubt the value of having systematic information on what students learn at school. Such information has obvious intuitive advantages, especially if it is available for different points in time, and thus allowing a judgment to be made about the stability, rise, or fall in standards. However, obtaining such information requires a data-gathering exercise, which in turn requires funding as well as careful decisions to ensure that the most appropriate information for policy makers and education managers is obtained.

While one can learn much from the experience of other countries that have carried out national assessments, it should be recognized that such experience represents considerable variation. For example, in deciding on which agency will carry out the assessment,

much will depend on a country's traditions, while the precise information that is obtained for indicators will depend on the country's experience in test development as well as the needs of policy makers. Countries that have long traditions of educational research may find that a national assessment fits quite readily into existing structures. Countries without such a tradition that are considering embarking on a national assessment, on the other hand, will have to give serious consideration to the steps involved in such an exercise and how they are to acquire the resources necessary to bring it to a successful conclusion.

For a country considering a national assessment that does not have a developed infrastructure for educational research involving the administration of large-scale surveys, an initial assessment should not be overambitious in the curriculum areas covered, assessment procedures, sample complexity, or demands on personnel. Almost inevitably there will be tension between the ideal of collecting as much information as possible and the need to use the initial exercise to provide basic data on the functioning of the education system and to develop local capacity. Keeping the scope of an assessment manageable—by, for example, limiting it to one curriculum area and one grade level—increases the chances of a successful operation. Another option is to limit its geographic coverage. Particularly in large, diverse countries, valuable experience and useful policy-related information can be obtained from assessments confined to one or a few regions of a country. However, if it is hoped to use data from national assessments to monitor achievement trends over time, limitations in the data-gathering procedures in the early stages will affect the ability to make comparisons in later years.

A first step for any country in the process of deciding whether or not to carry out a national assessment is to determine its cost. Only when one knows what costs will be incurred can one begin to consider whether the money required would be better spent on some other educational activity, such as improving school facilities or teacher training.

In making a decision about carrying out a national assessment it is also important to try to determine what kind of information from such an assessment is likely to be of use to policy makers, as well as to consider the way in which they might intervene in the education system in the light of the information revealed in the assessment. In this context, one may reflect on some of the following questions to which a national assessment might provide an answer:

- Is the overall performance of students in the education system in a particular curriculum area (e.g., reading, mathematics, science) at a particular grade or age level poorer or better than expected?
- Is the performance of students in particular domains of achievement (e.g., understanding of mathematical concepts) better or poorer than in other domains (e.g., ability to carry out mathematical computations)?
- Do we wish to obtain base-line data that will allow us to judge whether the performance of students in key curriculum areas (e.g., reading, mathematics) is improving or worsening over time?

- Is the performance of students in certain sectors of the education system less satisfactory than the performance of students in other sectors? This question can be asked about
  - –students in urban schools and students in rural schools
  - –students in public schools and students in private schools
  - –students from varying ethnic/language groups
  - –students in different geographical areas (provinces or states)
- Do student skills and knowledge differ by gender?
- Is there evidence of a relationship between students' knowledge and skills and such factors as the type of training that teachers have had, the amount of time they devote to a curriculum area in a class, or the amount of time that students spend doing homework?

An adequately designed national assessment can throw light on all these issues. It will not, however, normally point directly to specific solutions to any problems that it may uncover. In seeking solutions, it is likely that policy makers and managers will revisit frequently-tried solutions, choosing the one that seems most appropriate. These include improving the management of education, increasing the availability of textbooks, improving supervision of instruction, improving teacher training (both preservice and inservice), and reforming curricula (e.g., mandating that a greater amount of time be devoted to a particular curriculum area).

In considering approaches to educational reform on the basis of the findings of a national assessment, it is well to bear in mind that the choice of a policy maker or education manager in prescribing action to deal with deficiencies revealed in the assessment is largely constrained by a lack of knowledge of how the education system works. Although knowledge about schooling is only partial and does not approach an integrated theory of school organization, processes, and effects (see Bryk and Hermanson, 1993), there is a danger, given that national assessments are based on an input-output model of education, that prescriptions to deal with problems will also be based on that model. Such an approach would be short-sighted. It is important that individuals with responsibility for improving the educational service should look beyond simple input-output models and try to be more creative in their search for solutions, and to recognize that inadequate understanding in the past has often resulted in outside (government) efforts to control school inputs that fail to produce the intended consequences.

In considering reforms based on the data gathered from a national assessment, one should not lose sight of the fact that aspects of schooling that are not fully represented in the input-output model (or may even be misrepresented in it) may actually play an important role in affecting student achievement. For example, if we think of schools as systems, it becomes clear that the production function model of schooling, where independent cause and effect are dominant features, cannot adequately deal with the flow of information and communication involving a variety of feedback loops that is a feature of systems (Bryk and Hermanson, 1993). Furthermore, if it is true that communitarian aspects of schooling have important implications for student learning,

then again the input-output approach of our current system of indicators will not adequately represent what goes on in schools and, as a consequence, will be limited in its capability to diagnose problems or act as a basis for remedial action.

Such considerations have led a number of commentators to suggest a need for different types of indicators than the ones currently in use and for a different use for the information they provide. It has, for example, been suggested that an indicator system should be framed around six major topic areas that are of concern in education: learning outcomes, the quality of educational institutions, children's readiness to learn when they come to school, societal support for learning, the contribution of education to economic productivity, and equity. Further, it has been suggested that the data derived from indicators relating to these areas should be directed towards informing diverse audiences about the state of the education system, rather than being used as a basis for instrumental action. In line with the need to inform, work should be carried out to provide in-depth understanding of the forces behind the key indicators. To further this understanding, research, case studies, and program evaluations are recommended (Bryk and Hermanson, 1993; United States, National Center for Education Statistics, 1991).

It is difficult to distinguish these activities from the activities that have been the mainstay of educational research for a long time. If the message in the case for a different form of indicator is meant to imply that current work on monitoring involving indicators should be abandoned in favor of more traditional educational research, it is unlikely to be heard. First, traditional educational research has been modest in its contribution to meeting the needs of policy makers and education managers. Furthermore, the kind of program envisaged in the alternative would be extremely expensive. Rightly or wrongly, many policy makers see the current indicator movement as providing them with information that is more relevant than the information that traditional educational research provided and at a more affordable cost. They should not, however, ignore the concerns that lie behind the search for alternative indicators and an alternative way of responding to indicator information. The present system has at best serious limitations and at worst serious shortcomings that should not be ignored.

Whatever the reasons, the idea of a national assessment and its associated indicator system is becoming more popular across the world and it seems likely that it will be difficult in the immediate future for countries to resist the trend to document the outcomes of their educational endeavors. One may hope, however, that this will be done judiciously and that national assessments will not be carried out simply for the sake of carrying them out. Time spent in seriously considering the type of information that might be useful and how precisely it would be used would be time well spent before embarking on a national assessment.

REFERENCES

Baker, E.L., and Linn, R.L. (1995). United States. Pp. 197-209. *Performance Standards on Education: In Search of Quality.* Paris: OECD.

Bloom, B.S., Madaus, G.F., and Hastings, T. (1981). *Evaluation to Improve Student Learning.* New York: McGraw Hill.

Bottani, N., and Tuijnman, A. (1994). "International Education Indicators: Framework, Development and Interpretation." Pp. 21-35. *Making Education Count: Developing and Using International Indicators.* Paris: OECD.

Bryk, A. S., and Hermanson, K. L. (1993). "Educational Indicator Systems: Observations on Their Structure, Interpretation, and Use." *Review of Research in Education,* 19, 451-484.

Burnstein, L., Oakes, E., and Guiton, G. (1992). "Education Indicators." Pp. 409-418. M.C. Alkin (ed). *Encyclopedia of Educational Research* (6th ed). New York: Macmillan.

Canada. Council of Ministers of Education (1996). *School Achievement Indicators Program: Science Assessment Framework and Criteria.* Toronto, Ontario: Author.

Chinapah, V. (1992). *Monitoring and Surveying Learning Achievements: A Status Report.* Studies and Working Documents 1. Paris: UNESCO.

Chinapah, V. (1996). "After Jomtien: UNESCO's Current Policy on Assessment." Pp. 42-57. A. Little and A. Wolf (eds). *Assessment in Transition: Learning, Monitoring and Selection in International Perspective.* Oxford: Pergamon.

Elley, W. B. (1992). *How in the World Do Students Read? IEA Study of Reading Literacy.* The Hague: IEA.

Frith, D. S., and Macintosh, H. G. (1984). *A Teacher's Guide to Assessment.* Cheltenham, UK: IEA.

Gipps, C., and Murphy, P. (1994). *A Fair Test? Assessment, Achievement and Equality.* Buckingham, UK: Open University.

Goldstein, H. (1996). "International Comparisons of Student Achievement." Pp. 58-87. A. Little and A. Wolf (eds). *Assessment in Transition. Learning, Monitoring and Selection in International Perspective.* Oxford: Pergamon.

Greaney, V., and Kellaghan, T. (1996). *Monitoring the Learning Outcomes of Education Systems.* Washington, DC: World Bank.

Greaney, V., and Rojas, C. (1996). "Lessons Learned." Pp. 169-173. P. Murphy, V. Greaney, M. E.Lockheed, and C. Rojas (eds). *National Assessments: Testing the System.* Washington, DC: World Bank.

Great Britain. Department of Education and Science (1991). *Testing 7-year-olds in 1991: Results of the National Assessment in England.* London: Author.

Guskey, T. R. (1994). *High Stakes Performance Assessment: Perspectives on Kentucky's Educational Reform.* (2nd ed). Thousand Oaks, CA: Corwin Press.

Guthrie, J. W. (1991). "The World's New Political Economy is Politicizing Educational Evaluation." *Educational Evaluation and Policy Analysis,* 13, 309-321.

Himmel, E. (1996). "National Assessment in Chile." Pp. 111-128. P. Murphy, V. Greaney, M.E. Lockheed, and C. Rojas (eds). *National Assessments: Testing the System.* Washington DC: World Bank.

Ilon, L. (1996). "Considerations for Costing National Assessments." Pp. 69-88. P. Murphy, V. Greaney, M. E. Lockheed, and C. Rojas (eds). *National Assessments: Testing the System.* Washington, DC: World Bank.

Johnson, E.G. (1992). "The Design of the National Assessment of Educational Progress." *Journal of Educational Measurement*, 29, 95-110.

Kellaghan, T. (1996a). "IEA Studies and Educational Policy." *Assessment in Education*, 3, 143-160.

Kellaghan, T. (1996b). "National Assessment in England and Wales." Pp. 129-136. P. Murphy, V. Greaney, M. E. Lockheed, and C. Rojas (eds). *National Assessments: Testing the System.* Washington, DC: World Bank.

Kellaghan, T. and Grisay, A. (1995). "International Comparisons of Student Achievement: Problems and Prospects." Pp. 41-61. OECD, *Measuring What Students Learn.* Paris: OECD.

Kellaghan, T., Madaus, G.F., Airasian, P. W., and Fontes, P. J. (1976). "The Mathematical Attainments of Post-primary School Entrants." *Irish Journal of Education*, 10, 3-17.

Koeffler, S. (1991). *Assessment Design.* Paper Presented at the Seminar on Measurement/ Assessment Issues, Educational Testing Service, Princeton, NJ.

Lapointe, A. (1990). *Why a National Assessment?* Princeton, NJ: National Assessment of Educational Progress.

Lee, V.E., Bryk, A.S., and Smith, J.B. (1993). "The Organization of Effective Secondary Schools." *Review of Research in Education*, 19, 171-267.

Lockheed, M. E., and Hanushek, E. (1988). "Improving Educational Efficiency in Developing Countries: What Do We Know?" *Compare*, 18(1), 21-38.

Loxley, W. (1992). "Managing International Survey Research." *Prospects*, 22, 289-296.

Madaus, G.F., and Kellaghan, T. (1992). Curriculum Evaluation and Assessment. Pp. 119-154. P. W. Jackson (ed). *Handbook of Research on Curriculum.* New York: Macmillan.

Martin, M. O., and Kelly, D. L. (1996). *Third International Mathematics and Science Study. Technical Report. Vol 1: Design and Development.* Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Martin, M. O., Mullis, I. V. S., Beaton, A. E., Gonzalez, E. V., Smith, T. A., and Kelly, D. L. (1997). *Science Achievement in the Primary School Years. IEA's Third International Mathematics and Science Study.* Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Mehrens, W. A. (1992). "Using Performance Assessment for Accountability Purposes." *Educational Measurement: Issues and Practice*, 11(1), 3-9, 20.

Meisels, S. J., Dorfman, A., and Steele, D. (1995). "Equity and Excellence in Group-administered and Performance-based Assessments." Pp. 243-261. M. T. Nettles and A. L. Nettles (eds). *Equity and Excellence in Educational Testing and Assessment.* Boston: Kluwer.

Mullis, I. V. S., Martin, M. O., Beaton, A. E., Gonzalez, E. J., Kelly, D. L., and Smith, T. A. (1997). *Mathematics Achievement in the Primary School Years. IEA's Third International Mathematics and Science Study.* Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Nwana, O. C. (1996). "What Are National Assessments and Why Do Them?" Pp. 21-31. P. Murphy, V. Greaney, M. E. Lockheed, E. C., and Rojas (eds). *National Assessments. Testing the System.* Washington, DC: World Bank.

OECD (1995). *Education at a Glance.* Paris: OECD.

Olivares, J. (1996). "Inclusive National Testing: Chile's 'Quality of Education Assessment System.'" Pp. 118-133. A. Little and A. Wolf (eds). *Assessment in Transition. Learning, Monitoring and Selection in International Perspective.* Oxford: Pergamon.

Phillips, G. W., Mullis, V. S., Bourque, M. L., Williams, P. L., Hambleton, R. K., Owen, E. H., and Barton, P. E. (1993). *Interpreting NAEP Scales.* Washington, DC: National Center for Education Statistics.

Rojas, C. (1996). "The Colombian Education Assessment System." Pp. 147-155. P. Murphy, V. Greaney, M. E. Lockheed, and C. Rojas (Eds.), *National Assessments. Testing the System.* Washington, DC: World Bank.

Ross, K. N. (1987). "Sample Design." *International Journal of Educational Resarch,* 11, 57-75.

UNESCO (1990). *World Declaration on Education for All. Meeting Basic Learning Needs.* New York: UNESCO.

United States National Center for Education Statistics. (1991). *Education Counts: An Indicator System to Monitor the Nation's Educational Health.* Washington, DC: USNCES.

Weiss, C. H. (1979). "The Many Meanings of Research Utilization." *Public Administration Review,* 39, 426-431.

Woodhouse, G., and Goldstein, G. (1996). "The Statistical Analysis of Institution-based Data." Pp. 135-144. H. Goldstein and T. Lewis (eds). *Assessment: Problems, Developments and Statistical Issues.* Chichester, UK: Wiley.

CHAPTER 3

EVALUATION AND CURRICULUM STANDARDS: INDICATOR SYSTEMS IN AN ERA OF EDUCATIONAL REFORM

Gilbert A. Valverde

The preceding chapter addressed the issue of national-scale evaluation of the results of academic learning. The following chapter provides a bridge between national aspirations or educational goals and the results obtained through the monitoring and evaluation education processes. Such processes focus on the school curriculum, i.e., on what is being taught and how it is being taught. The author discusses the characteristics and advantages of a systemic model and illustrates the discussion with information and methodologies developed in conjunction with the Third International Mathematics and Science Study, or TIMSS. The chapter concludes with an analysis of the implications of this perspective for educational policy in domains such as school textbooks, curriculum, indicators of educational development, and reform strategies.

INTRODUCTION

Educational reform is sweeping the Americas. A number of nations have embarked on a variety of policies devoted to the profound reform of existing educational systems (Álvarez, 1997). It is particularly important to note the *systemic* aspect of many educational reform movements in the region, as it has profound implications for educational policy making in general, and especially for the design of evaluation and monitoring systems.

The contemporary educational reform movement defends the point of view that needed reforms require comprehensive policies that are directed at many different aspects of the educational system. Certainly this point of view has been defended by programs devoted to promoting educational reform in Latin America (see Slavin, 1994), and it is documented in various descriptions of reforms currently underway throughout the region (e.g., Álvarez and Ruiz-Casares, 1997). Systemic policies—that is, policies intended to rearrange authority and resources among individuals and agencies in order to alter the system by which education is delivered—present a formidable challenge to policy makers and evaluators alike.

Understanding that structural features of an educational system are central to reform efforts and the effects of these efforts is vital to the future of educational reform policy. An appreciation of how educational policy can alter these structures, and what the probable results will be, is greatly enhanced through considering educational goals, processes, and products as an integrated system. In short, to better understand how to change or reform education, policy makers must have at their disposal indicators that measure education *systemically.*

A meaningful measurement of educational systems requires a comprehensive conceptual framework and a corresponding array of indicators designed to relate the various parts of the system to each other, and to their outcomes. If the technologies for the evaluation of educational reform are to move forward, we need methodologies to collect data rigorously, consistently, and on a significant number of factors associated with the *processes* of schooling—not merely on goals and products. This makes meaningful analysis of policy alternatives possible, resulting in a deeper understanding of existing systems as well as potential alternatives.

Considering indicator systems in light of their potential for illuminating policy is clearly critical at this time for many Latin American countries. Current policy enacted or being considered in the region include "top-down" policies of setting goals from a centralized entity (most often a ministry or similar provincial/regional authority) or "bottom-up" strategies intended to foster pedagogical innovation at the school or classroom level. Frequently, both types of policies coexist in recognition of the importance of pedagogical innovation in revitalizing educational delivery systems in a context in which national or regional authorities are still required to ensure minimum standards in these delivery systems.

Whether a country is enacting "bottom-up" or "top-down" policies—or a mix of both—indicator systems play an important role. However, their importance is directly related to whether or not they provide information that enables policy makers to assess the relative merits of alternative policies. Currently, most educational policy making is intended to increase, among other things, the efficiency and quality of education. Such a context requires indicators that not only assess what children learn, but how they are taught and the types of results achieved. Thus, not only do policy makers require indicators of student achievement, they also require indicators of the processes that explain achievement in order to assess options that have the potential for furthering their goals.

This paper will first review how monitoring policies regarding both educational goals (such as curriculum standards) and educational products (measures of student achievement) inevitably lead to an evaluation that considers the importance of the *processes* that transform goals into products. A case is then presented for envisioning new models for evaluation that make the relationship between these two policy dimensions the explicit focus of the evaluation process. Educational processes then become an integral link between goals and outcomes that must be accounted for in indicator systems.

Some of the essential characteristics of a systemic model will then be addressed. This is the systemic model of educational opportunities resulting from the work of the Survey of Mathematics and Science Opportunity (SMSO) at Michigan State University for the recently completed Third International Mathematics and Science Study (TIMSS) conducted by the International Association for the Evaluation of Educational Achievement (IEA). These characteristics are intended to guide the development of *integrated* evaluations of curriculum, instructional goals and practices, and student achievement that constitutes an innovation in the field of educational evaluation.

The paper will delineate how the study of curricular components can provide policy makers with valuable indicators in a climate of educational reform. It will also explain how educational systems are organized to deliver the curriculum to students. This paper will explore these methodological innovations, detail their contributions to the evaluation of educational policy, and provide examples from data collected for studies conducted by the SMSO and the U.S. National Research Center for the TIMSS.

EDUCATIONAL POLICIES, INDICATOR SYSTEMS, AND THE INVISIBLE PROCESS

Since the early 1980s, the nations of America have embarked on the formulation of policies intended to address perceived shortcomings of current educational delivery systems. Two important results of these policy initiatives and associated evaluation strategies have been the development of assessment systems and the reformulation of existing policies of curricular governance, or the formulations of new policies in this area. Policies regarding assessment and goals are often conceived of as joint mechanisms of system control or "system management" (Apple, 1990) in which measuring products (most frequently through achievement tests) against purposes (curriculum standards) is held to provide a scientifically valid, and politically sufficient, evaluation.

However, an increasingly large body of compelling evidence has been collected from research on assessment and curricular governance regarding problems in both testing and curriculum policy that have important implications for reform-oriented policy making.

Pitfalls in assessment and content-driven reform

For more than three decades many countries around the world have conducted a variety of cross-national, national, and subnational assessments of educational achievement. All have claimed, to a greater or lesser degree, to have provided important data useful for ascertaining the effectiveness of educational systems. Educational policy in Latin America reflects these international trends. This is evident, for example, in a recent description of evaluation in a Colombian curricular program proposal: The evaluation will search for failures and successes in order to include the necessary corrective measures that would guarantee the progress of the student (Ministerio de Educación Nacional de Colombia, 1990).

In many contemporary systems, a primary (if not sole) role of evaluation in the examination of the effectiveness of programs and policies is the use of student achievement tests. The implicit assumption of such approaches to evaluation is that the student achievement scores reported in such assessments can, in fact, be attributed to students' educational experiences, and therefore represent a valid assessment of the comparative effectiveness of educational programs and policies.

Increasingly, it has been argued that assessments, particularly national assessments, can be primary motors of educational reform. Certainly this is the case in recent debates in the United States, including President Clinton's call for a national testing program. It is also apparent in the expansion of Latin American efforts in testing. The implicit deductive scheme leading to such policies appears to have the following components:

- High-stakes tests will clarify and make explicit the goals of education.
- They measure whether teachers, schools and students successfully arrive at these goals.
- Holding teachers, schools and students accountable through the use of these assessments will provide them with the necessary motivation to improve teaching and learning.

It has been claimed that such a deductive scheme is based on behaviorist psychology and pedagogy (Noble and Smith, 1994) and is thus inconsistent with most current reform efforts in the Americas. Many reforms promulgate the use of modern pedagogical approaches pursuing the teaching of higher-order thinking skills and critical thought, some even recommending the use of constructivist pedagogies (Ministerio de Educación Nacional de Colombia, 1990; Ministerio de Educación Pública de Costa Rica, 1996; Secretaría de Educación Pública de México, 1993).

Closed-system assessment programs, by focusing on matching goals to outcomes, ignore the educational *process*. They implicitly assume that the process of schooling will "take care of itself" once goals are clarified in tests. In effect, these assessment regimes paradoxically ignore *schooling*—that is the process whereby instructional goals are implemented in the classroom. They hold teachers, students, and schools accountable for outcomes, disregarding the question of whether they have any control over the factors that provide children the opportunity to learn assessed skills.

A recent review of many studies of testing reveals that such policies affect students and teachers in ways incompatible with current reform movements. It documents that such evaluation policies have created environments in which teachers ignore topics that the tests fail to cover, neglect team-teaching approaches, emphasize basic-skills over higher-order thinking skills—and even match the format of their teaching to the format of the tests (Noble and Smith, 1994).

We are therefore confronted with the vital question of whether it is possible for such assessments to identify faults and strengths *related to the educational experiences of children* that would make it possible to identify corrective measures.

At the heart of this challenge lies the question of whether or not these assessments do in fact measure what they claim (Airasian and Madaus, 1983). The problem confronted is one of the *instructional* and the *curricular* fairness of the assessments. Simply put, it is the problem of whether tests measure the objectives of the curriculum, and whether schools provide learners with instruction in the skills and knowledge assessed. The answer to this problem represents a critical element in determining the policy relevance of assessments and lies at the very core of the use of assessments in educational policy.

IEA studies introduced the notion of opportunity to learn (OTL) as a means of ensuring the technical validity of their findings (McDonnell, 1995), and by doing so, introduced not only a concept that revolutionized the technology of educational assessment, but also provided a tool to enhance the policy relevance of assessments.

OTL provided the IEA studies a measure of how close a match existed between the tests administered and educational practices in each of the participating countries. Initially, teachers were requested to look at a copy of the items on the assessments and report whether they had provided their students the instruction necessary to solve these items correctly. The simplicity of these early measures belies the enormous conceptual importance of OTL—it was a first step in recognizing the imperative of characterizing *instruction* in order to explain *achievement*. OTL was the measure of opportunities necessary to perform well on achievement tests, and is thus a recognition that the instructional practices of teachers contribute to the attainment of students.

In essence, OTL is the measure of process. Traditional assessments were unable to measure the variety of instructional experiences that explains the knowledge and skills they measured. OTL instruments began as measures of the content covered by the teachers of children participating in assessments. Even in the most centrally governed educational systems, there is a wide variation in the topics and skills covered by teachers. Researchers have shown that the opportunity to learn the skills being tested is a significant explanatory variable of student performance (Burstein, 1993; Burstein et al., 1990; McDonnell, 1995; Muthén et al., 1995). Since the early appearance of OTL as a measure of content covered in class, this concept and the attendant technologies of measurement have become more complex and more capable of portraying the variation in educational opportunities provided to children (Guiton and Oakes, 1995; McDonnell, 1995; Schmidt et al., 1996; Schmidt et al., 1997a; Shavelson and Webb, 1995).

OTL provides the vital link between goals and outcomes that turns mere testing into more solid educational evaluation. *It recognizes that how educational resources are used is every bit as important as which resources are used.* Including OTL measures in educational assessments provides indicators on the educational process as it unfolds in schools and classrooms, and provides policy makers with vital tools for the evaluation and formulation of reform policies. This preoccupation with goals and processes is extremely important given the particular character of current educational reform policies.

A major set of research and policy initiatives aimed at improving education in the Americas is the movement for content-driven systemic school reform. A considerable

body of work in the United States has contributed to supporting this type of reform (see Clune, 1993; O'Day and Smith, 1993). The central tenet of these initiatives is that the reform of educational systems must be driven by *curriculum*. It is stated that ambitious curricula must be formulated and that appropriate mechanisms must be designed to implement these curricula so that students may have the opportunity to attain high levels of achievement. Content-based reform holds that a core curriculum provides a basis for determining which resources are necessary to ensure that students are provided the opportunities required to master them. Thus, the curriculum would directly impact teacher training and certification, school course offerings, instructional resources, and structures of educational governance and accountability. In fact, it is sometimes held that contrasting the curriculum with educational practices in the classroom is *required* to accurately monitor how students are provided with adequate opportunity to learn, and is one of the evaluation responsibilities of a reformed system.

High expectations concerning the role of curriculum standards are certainly held in Latin American countries; a case in point are the statements made by the current president of Mexico, while serving as Secretary of Education: "The planning of programs of study is one method for improving the quality of education" (Zedillo Ponce de León, 1993).

Theoreticians of content-driven systemic reform make reference to educational systems with strong national curricula in formulating their policy proposals. For example, O'Day and Smith (1993) state, for the case of the United States, that: "When fully implemented, this model of content-driven systemic reform would be a uniquely American adaptation of the educational policies and structures of many of the world's highly developed nations" (p. 252). Clune (1993) also mentions that a possible result of content-driven systemic school reform would be a centralized system of governance and curricula "resembling the national curricula and educational ministries of countries that have high levels of student achievement" (p. 233).

Problems arise from a number of instances in which *standards* are only formulated as pedagogical goals for the system, without regard to the actual pedagogy that leads to their realization. When such standards are coupled in *closed system* assessment programs as outlined above, the consequences for reform can be disastrous.

OTL, as we have seen, arose as an answer to a very basic question regarding the fairness (which, for assessment technicians, is called *validity*) of assessments. Is it fair to judge students against performance on tests for which their educational experiences have not prepared them? Is it fair to judge teachers or schools against the performance of their

students on such assessments when they have no control over the instructional resources necessary to teach the requisite skills? An educational policy analog to OTL also exists. These are standards concerning the provision of opportunities to learn.

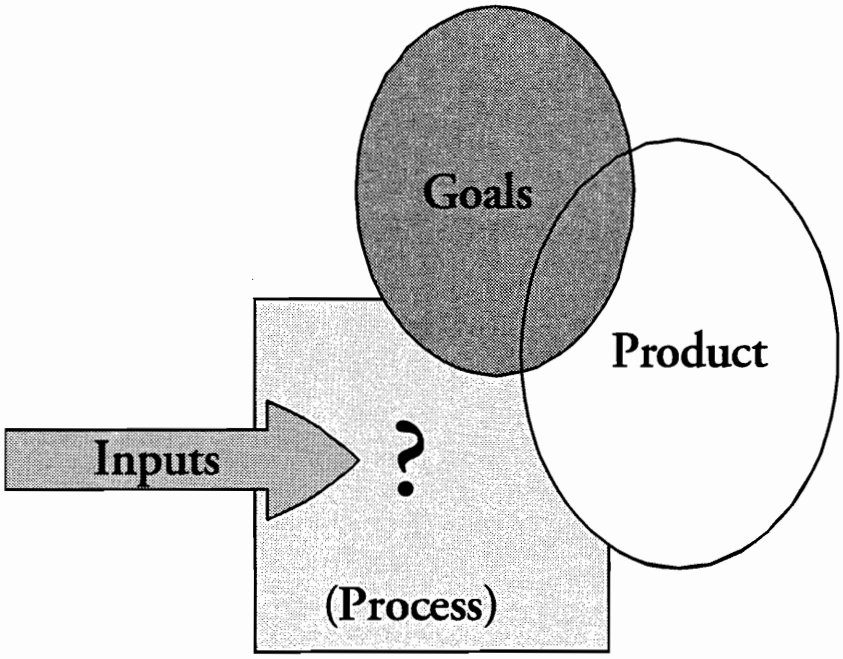
Sometimes known as *school delivery standards* (McDonnell, 1995; Porter, 1993), they have the potential of enhancing the effectiveness of curriculum policy by performing two very important functions (Porter, 1993), one symbolic, the other technical.
The symbolic function of school delivery standards can be to provide a new image or paradigm, a sense of "imaginative possibility" (Apple, 1992) concerning what ideal instruction might be. Well formulated, such images may motivate teachers to be innovative and creative. The technical function relates to the monitoring of implementation. School delivery standards can also provide the basis for regulating the actions of teachers, school officials, and other educational and political agencies.

Contemporary reform strategies include a press toward teaching less simplistic, factual, or algorithmic knowledge and more conceptual and critical skills. Clearly, such an emphasis requires, in the case of many educational systems, shifts not only in *what* is taught, but also in *how* it is taught. This is of course a primary inducement for policy makers to consider both the setting of curricular goals and the establishment of delivery standards.

A central tenet of advocates of systemic educational policy is that curriculum must drive it. These advocates hold that ambitious curricula must be formulated and that appropriate mechanisms must be designed to implement these curricula so that students may have the opportunity to attain high levels of achievement. Content-based reform holds that a core curriculum provides a basis for setting school delivery standards. Thus, the curriculum would directly influence teacher training and certification, school course offerings, and instructional resources and structures of educational governance and accountability (including test design and administration). In fact, it is held that contrasting the curriculum with educational practices in the classroom is required to accurately monitor how students are provided with adequate opportunity to learn, and is an evaluation responsibility of a reformed system.

Thus, contemporary reform strategies, as well as assessment technologies, require that the processes of education be examined much more closely in ways that enhance the usefulness of the information for serving reform strategies. The dilemma, as presented in Figure 1, is to devise models of evaluation that unpack the "black-box" of the educational processes in order to understand how inputs and goals are transformed into products.

Figure 1. The Problem of Integrating Process into Evaluation Strategies



Note: The challenge for evaluation is to account for processes in ways that directly link them to goals, inputs, and products.

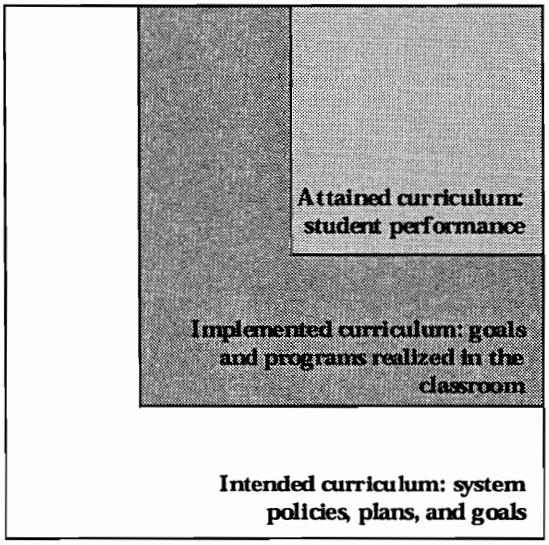
THE VISIBLE PROCESS: A SYSTEMIC MODEL OF EDUCATIONAL OPPORTUNITIES

In an era of educational reform, educational policy makers require policy-relevant evaluation that can uncover the institutional and pedagogical correlates of differences in achievement levels. Policy makers and researchers alike are concluding that the proper role of educational evaluation is to help uncover the systemic features, instructional practices, and other educational inputs and processes that are best suited to shape desired educational outcomes.

In designing data collection and analysis strategies for the TIMSS, a study of educational systems as they currently exist to deliver educational experiences to students was used. A complex array of data collections interrelated through a systemic model of educational opportunities, within which the achievement data are but one important component, was contrived. We think of these educational opportunities as the formulation of national or subnational intentions and their classroom implementation. They are policy instruments and teaching techniques intended to guide realized experiences. The model recognizes that actual learning or realized educational experiences are subjective processes realized by each student.

The starting point for model development is a tripartite model of curriculum, as shown in Figure 2, which has been a traditional feature of IEA studies for some time. This model makes an analytical distinction between curriculum as system goals, curriculum as instruction, and curriculum as student achievement. Each of these dimensions is known, respectively, as the intended, implemented, and attained curriculum.

Figure 2. IEA Analytic Model for Curriculum



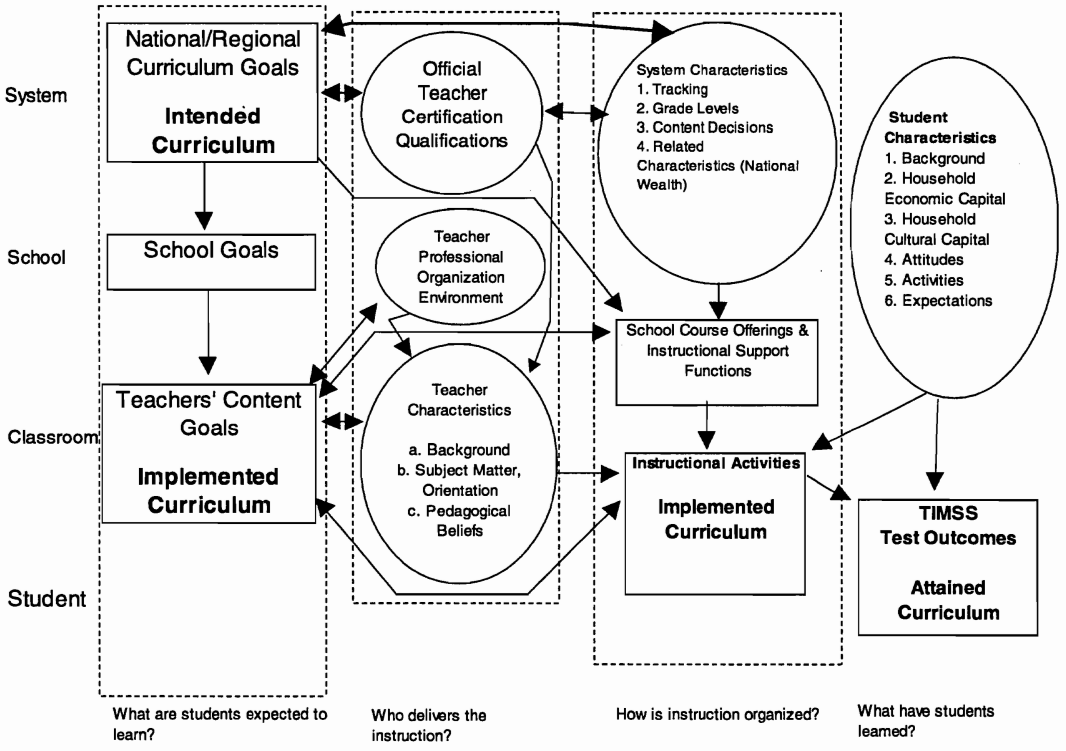
The systemic model of educational opportunity, developed at the U.S. TIMSS Center through a separate grant known as the Survey of Mathematics and Science Opportunities (SMSO)[1], centers on the premise that one purpose of an educational system, and a primary focus of its pedagogical role, is to enable educators and policy makers to provide opportunities to learn the sciences and mathematics. In this schema we recognize curricula, schools, and teachers as elements that both define and delimit the potential learning experiences afforded to students for learning mathematics and the sciences. The opportunities to reach a learning goal (e.g., have useful knowledge and skills centered around specific topics) have a direct impact on the actual accomplishment of that goal by students.

The term *goal* is used to refer to content as defined by analytic curriculum frameworks (Robitaille et al., 1993; Survey of Mathematics and Science Opportunities, 1992a; Survey of Mathematics and Science Opportunities, 1992b), and includes performance expectations and attitudes. This section presents the salient features of the SMSO systemic model of educational opportunities that is the theoretical framework for the TIMSS data collections (Schmidt et al., 1996; Schmidt and McKnight, 1995). This model is organized around four fundamental questions: What are students expected to

learn?; Who delivers the instruction?; How is the instruction organized?; and What have students learned?

Figure 3 depicts the constructs and interrelationships in the model. The four columns represent the four primary research questions, and the four rows specify the four levels of the educational system to be examined. The arrows show the network of relationships among the constructs.

Figure 3. SMSO Systemic Model of Educational Opportunity



Note: This model recovers the educational process through explicit linkages to inputs, goals, and products. These linkages are made through the category system of the SMSO analytical curriculum frameworks.

What are students expected to learn?

Describing educational opportunity begins with the description of the knowledge and skills students are supposed to master. There are three main levels of the educational system at which goals are commonly set: the national or regional level, the school site level, and the classroom level.

The first question involves not only the determination of learning goals for a system or country as a whole, but the differentiation of such goals for subdivisions of the educational systems, including regions, tracks, different types of schools, and classrooms. Targeting each level of the educational system, the general questions become:

- What goals are pursued at the national or subnational level?
  a. How are these goals aligned to grades and how are they sequenced?
  b. Who decides which goals are pursued and how are these decisions are made?
  c. How, and on what basis, are pupils differentiated with respect to the goals they are to pursue?
- What goals are pursued at the school level (or for various types of schools), and how are these influenced by national or subnational goals?
- What goals are pursued at the classroom level or for various types of classrooms of students? What influence is exerted by national, subnational, or school-wide curriculum objectives?

The substantive issues here concern the curriculum. In the parlance of IEA, the learning goals specified at the national/subnational level is the intended curriculum, while the specification of learning goals at the school or classroom level is at least part of what is designated as the implemented curriculum. The curriculum, as goals and plans for the distribution of potential learning experiences, is one of the major concepts in this model. As suggested by the preceding questions, the TIMSS studies the concept at each level of the educational system.

Who delivers the instruction?

Educational opportunity is also characterized through the delivery of instruction, which is the central role of teachers. One way to characterize a teacher's role is to describe the official teacher certification qualifications, which may include grade and subject restriction, educational attainment requirements for awarding each license, type of educational degree required, and any specific course work or practical experiences that may be required.

The teachers' professional environment also influences the delivery of the instruction. Environment refers to such things as time usage, including the proportion of professional time spent in actual teaching or in planning coursework.

Since teachers play such a central role in the education system, portraying them in detail is critical. Many studies have shown how teacher characteristics influence quality of instruction, and hence the quality of the educational opportunity (Carlsen, 1991; Cohen, 1988). Two broad categories of a teacher's characteristics are especially important: background and beliefs.

Teacher background variables include age, gender, education, subject taught, and teaching experience. Teacher beliefs address subject matter orientation, and subject matter-specific pedagogy. Teachers' beliefs about subject matter (the views they hold

about the disciplines they teach) can affect instructional practices and student achievement. Teachers' pedagogical beliefs, on the other hand, refer to their notions about the best way to teach a particular topic within a discipline.

Underlying the research for this section is an examination of teachers' personal characteristics: what they believe about school subjects and how they are best taught, and how they are trained. As all these characteristics influence the delivery of curricula—or of potential learning experiences—and they allow us insight into the types of alignment that exist between teacher training and practice, and national/subnational curriculum objectives. This data, when coupled with achievement data, can test the assumption of systemic reform advocates that such alignments affect educational attainment. It will also help identify particular subject matter and pedagogical beliefs that will best ensure the delivery of challenging curricula to students throughout the system.

How is the instruction organized?

Along with expectations for student learning and the delivery of instruction by teachers, the manner in which the instruction is organized influences the implemented curriculum, and therefore, the potential learning experiences of students. No single decision maker determines the organization of instruction. The array of possible loci of decision making concerning instruction is wide. Decisions may be made at the very top of the educational hierarchy or at intermediate levels. The locus of decision making may also be at the school site—with school administrators or with the classroom teachers charged with deciding how they teach content to their students.

This diffusion of decision-making authority is manifested in many ways. For example, there are important variations in the age-grade structure of educational systems and in the nature of the schools that serve different arrays of grades. Also, the school system often organizes students into different curricular tracks. There is no doubt that these circumstances determine organizational characteristics of the educational system. They also influence the qualifications of the teaching force and the types of instructional resources available to those teachers. In addition, they have a great bearing on the time and material resources available to students.

Implicit to how the instruction is organized are the school course offerings. The functions that schools perform that support the offering of courses in various school subjects, and the roles that individuals in the school perform in support of those functions, affect the instructional organization.

The organization of instruction also occurs in individual classrooms. Thus, the factors influencing the implemented curriculum include such elements as: textbook use, structure of lessons, uses of instructional material, student evaluation, student participation, homework, and student in-class grouping.

The organization of national and subnational goals into programs of study and into course offerings are decisions frequently made at the school site. Therefore, school

authorities are important potential sources of information on course offerings. Teachers can provide extensive data on their instructional practices and on how they organize lessons and deliver them to students.

What have students learned?

With the purpose of defining the educational system being to identify the factors that determine what and how students learn, examining student characteristics is important. Beyond the influence of curriculum goals, teachers, and instructional organization, student characteristics influence the process by which potential learning experiences are actualized. Schools cannot provide actual learning experiences, only opportunities for such experiences. In the end, the interest, motivations, aspirations and other characteristics of students, influenced by their personal backgrounds, transform potential experiences into realized experiences. Important student characteristics include the following: the student's academic history, the economic situation of the student's family and the student's socioeconomic status, cultural involvement of family, self concept, time spent outside school, motivation, and interest.

Identifying and measuring every component of an educational system is neither possible nor desirable. However, this model of educational experiences recognizes the interconnections between the elements of the educational system in a way analogous to the conceptualizations of the proponents of the systemic reform movement in the United States (Clune, 1993; O'Day and Smith, 1993). This is a generic model that can be used to describe many specific educational systems. It does not advocate a particular system, but is rather intended as a template against which to identify different systemic variations.

A STRATEGY FOR SYSTEMIC EVALUATION

Having presented the model of potential educational experiences, I will now discuss a measurement strategy, also developed by the SMSO, that links the model to various data collections. This strategy, developed for the cross-national study of mathematics and science education, was intentionally designed to apply to national evaluations, as it was the conviction of the research group from the outset that *cross-national evaluations* are only useful if they are also useful *national evaluations*.

The subdivision of curriculum into three components (see Figure 1) is intended as an analytical tool, a way of isolating aspects of curriculum for study. However, the TIMSS data collections are designed in an integrated fashion. The integrating elements are the TIMSS curriculum frameworks.

The TIMSS directly links the curriculum and subject matter elements of the model through the curriculum frameworks. The frameworks provide the unifying language for the TIMSS as a whole and for the curriculum analysis component in particular. These frameworks represent a multicategory, multiaspect specification of the content and performance expectations and perspectives in a form relevant to the measurement of

curricular material and tests. They provide a set of conventions that serve as the common language system. In brief, they represent the operation of mathematics and science as measurable entities in the provision of opportunities to learn.

The TIMSS curriculum frameworks are used to characterize all components of the TIMSS data collections. Courses, test items, achievement scales, textbooks, curriculum guides lessons, and teachers' pedagogical beliefs are all depicted in terms of framework categories and can be linked throughout the analyses. For example, we can link curriculum coverage in curriculum guides and textbooks on the subject of "estimating computations" (one category from the content aspect of the TIMSS mathematics frameworks) with student performance on a set of achievement items covering this topic. We also see how the performance expectations required to solve the particular set of items, which could be, for example, "selecting or constructing mathematically equivalent objects" (a category from the performance expectation aspect of the framework) is linked to the presence of such performance expectations in textbooks or curriculum guides. We can also link teacher responses to pedagogical beliefs about teaching "proportionality" with actual lessons on proportionality, and with a textbook presentation of that content.

The use of these types of analytic frameworks to tie together every data collection instrument provides a unique opportunity to study the alignment of all elements of the system in providing opportunities to learn defined curricular elements. These frameworks are the central technology designed to link measures of intention, implementation, and achievement.

The measurement of intention

The SMSO systemic model of educational opportunity recognizes the importance of national or subnational specifications of learning goals that are understood to embody the intended curriculum. Such statements of student learning goals are often found in official documents such as curriculum guides and programs of study. These documents, however, vary in the detail with which they specify the learning goals. Although not always official, textbooks also provide information on student learning goals. The information is usually detailed and combines information about goals at the official or semi-official level with information on the likelihood of various goals at the school and teacher level.

The curriculum, by specifying the learning goals at the national or subnational level, sets parameters that emphasize certain potential learning experiences and constrain others. For example, in a country with a mandatory national curriculum promoted by national assessments and a school inspectorate, the inclusion of a learning goal does not guarantee that it will be covered—that is, that the opportunity will actually be provided in classrooms—but it does greatly increase the probability of that event. The absence of a goal similarly increases the probability that potential learning experiences related to that goal will not be provided, but as before we deal only with changes in probabilities—in the probability distributions of potential learning experiences—and not with certainty that opportunity will not be delivered.

Differences across provinces, municipalities, or schools in the specification of learning goals, and the policies related to the learning goals, are critically important in understanding these relationships. The system-level specification of learning goals sets parameters by which potential learning experiences are constrained—although perhaps not in equal degree—no matter what the type of system.

The model envisions the curriculum as a primary defining element of potential educational experiences. It shapes national goals and expectations for learning. These potential experiences are themselves important systemic features and are closely tied to educational policy and issues of educational quality. Characterizing these features provides an understanding of the types of philosophies and goals that underlie them, and provides a fundamental element to understanding the context within which curricula are implemented.

Accordingly, curriculum is an outcome of equal importance to student achievement. How the intended curriculum is designed, what it includes, what policies of control are associated with it, and who decides what students are expected to learn are all fundamental political features of educational systems. These features are of central concern in any formulation of reform policies. The ability to link these features to other outcomes such as instructional practices and student achievement is the other strength of this model: it permits us to study how potential educational experiences become transformed and actualized as the curriculum is implemented in schools and classrooms.

The aims of TIMSS include analyzing the relationships of concepts, shown in Figure 1, that lie outside of achievement, such as national/subnational goals. These goals set the parameters or constraints within which potential learning experiences are delivered. Thus, an understanding of those constraints is central to improving the provision of educational opportunity. Policy makers, curriculum developers, and educators in many countries are struggling with the reform of their curricula. What should a "world-class curriculum," for example, in mathematics and the sciences, have as its content and as its performance expectations? Should it include noncognitive goals as well? What attendant policies need to be associated with such a curriculum? If policy makers are serious in wanting to address these issues as they undertake reform, then comprehensive and document-based descriptions of what different countries are actually doing are necessary to begin such deliberations.

The first purpose of the SMSO curriculum analysis is to answer, at the system level, the question: What are students expected to learn in mathematics and the sciences? As described earlier, the specification of national/subnational intent or goals delineates the defining characteristics of educational opportunity at the broadest level. The inclusion or exclusion of a specific content or performance in the national/subnational goals affects the probabilities associated with the provision of such potential learning experiences. In this way, the intended curriculum provides the necessary—but only in a stochastic sense—condition for the provision of educational opportunity. The study of such parameters provides the fundamental foundation from which a country's opportunity structure is interpretable.

The important links of the intended and the implemented curriculum serve as an example. Examination of these links is equivalent to the exploration of the structure of educational opportunity. What these links look like, what role curriculum guides and textbooks play, and how policies such as national systems of control, grade-level organization, and tracking affect the links, are all important questions in the study of potential educational experiences.

Additionally, from the point of view of national ministries or subnational authorities, a characterization from an international perspective of national/subnational intent along with the attendant policies is needed to inform policy decisions. What the national/subnational goals are, and what policies of enforcement are associated with them, are matters for national or subnational policy setting. No longer do most policy makers wish to set such policies in isolation, or in ignorance of what others—particularly their economic competitors—are doing.

From these two points of view—a study of educational opportunity and curricular policy setting at the ministry or subnational level—the characterization of the curriculum, as an answer to the question of what students are expected to learn, is essential. Yet so also are the attendant questions of who determines learning goals in a country, in what form and with what document structures the goals are specified, what policies exist in support of the learning goals, and in what ways these factors of the educational system are interrelated.

The question taken from the SMSO systemic model of educational opportunities with which this the first aspect of the curriculum analysis is concerned is: What are students expected to learn in mathematics and the science at the national or subnational level?

The measurement of implementation

Schools and teachers, through their characteristics and activities, also help to frame the potential learning experiences provided to students. The curricular organization of the system and the school, and the characteristics and subject matter knowledge of the teachers affect the provision and quality of potential learning experiences. The learning goals used by the teacher in the classroom, and the course offerings provided by the school further delimit and shape those opportunities. Thus, the next step in understanding educational opportunity is to measure what happens in the delivery of instruction to students. *Specifically, what are the concrete measures taken by people or organizations to ensure the actual fulfillment of curriculum policy?*

Implementing the curriculum entails the full scope of operations performed by the teachers and school administrators—those people entrusted with carrying out educational policy. Thus, school goals, teacher content goals, opportunities to learn, and instructional practices are all important aspects of implementation. Other important characteristics are the teaching force, the schools, and the national/subnational system that heavily influence the quality of implementation.

The SMSO has designed instrumentation to measure teachers' content goals, school goals, and more classic measures of OTL. For teacher content goals, the instrument measures what teachers intend to teach students and how many periods are allocated to the instruction of the content. The methods measure content goals against the TIMSS curriculum frameworks and link them with other aspects of implementation and with the measures of the intended and the attained curriculum.

The OTL aspect of the instrumentation has been substantially redesigned in respect to previous IEA studies, and is linked through the curriculum frameworks to all other measures. The redesign entails identifying groups of items from the achievement tests as exemplars of different content categories from the TIMSS curriculum frameworks. Teachers are asked to respond whether they have taught or intend to teach the content represented by the items and presented in terms of the frameworks. This new design is intended to increase the likelihood that teachers will provide OTL data that will not center on the specifics of individual items, but rather on the content categories that they represent.

There are also qualitative aspects of implementation. The personal characteristics of those people charged with the responsibility for realizing national intentions in the classroom mediate implementation. Characteristics of the schools and the educational system within which implementation takes place also influence implementation. A variety of measures survey the criteria that potential teachers must meet to be officially qualified to practice as educators and to practice in a particular setting (types of schools, working with certain types of students, etc.). Other portions of the instrumentation measure the professional environment that teachers experience in their schools, and explore a variety of personal characteristics of teachers. These characteristics especially include their subject matter orientations and pedagogical beliefs regarding the teaching of school subjects.

The delivery of instructional activities and the provision of potential learning experiences mostly takes place in the local school and, more precisely, in the classroom under the guidance of an individual teacher. Because of the "practical" character of these activities and the limited ability of most systems to monitor school or individual teacher decisions, major variations in instruction exist across schools and across classrooms within schools, even within "centralized" systems. Consequently, a primary focus for studies of potential learning experiences must be instructional practices—the actual actions carried out in the classroom.

In contrast, since goals are intentions, not provided experiences, they may be set at any or all levels of a system. Thus, a "centralized" system might be one in which national goals are specified and in which great efforts are made to control classroom activities, or it might be one in which such goals exist, but autonomy is permitted at many levels in the delivery of instruction. Distinguishing goals from actual activities and experiences is extremely important at both the conceptual and empirical levels. Therefore, in looking into how instruction is organized, the TIMSS instrumentation[2] explores the following sets of questions:

1. How is the system at the national or subnational level organized?
   a. What grade levels does the system have?
   b. How are the grades grouped into schools (primary, lower secondary, upper secondary, etc.)?
2. What roles and functions do schools play in the delivery of instruction, and how are mathematics and the sciences organized in school offerings?
   a. What are the programs of study?
   b. Are the subjects taught sequentially or simultaneously?
   c. What roles do schools play in supporting instruction (e.g., resource allocation vs. community, student, and family relations)?
   d. How are resources allocated to instruction?
   e. How is this related to teacher learning goals?
3. How is classroom instruction organized?
   a. How are lessons structured?
   b. How are textbooks and curriculum guides used in instruction?
   c. What instructional activities are used, such as laboratories, lectures, demonstrations, groupings, student evaluations, etc.?
   d. How are in-class practice and homework used in instruction?
   e. In what ways do students participate in the instruction?
   f. How is this related to teacher learning goals?

The organization of instruction occurs in both the school and in individual classrooms. The system itself has a structure by which instruction in a broader sense is organized. The latter has profound implications for instructional organization at the school and classroom levels, and together they play a major role in determining the quality of potential learning experiences.

The measurement of attainment

In the systemic model of educational opportunities, the measurement of achievement entails the measurement not only of what mastery pupils have attained, but also of other student characteristics. Just as it is necessary to distinguish between intention (or potential experience) and implementation (activities related to the further definition of potential experiences)—it is also necessary to distinguish between potential experiences and the active engagement of a student in an experience, and attainment (the changes produced within the student, as reflected on tests, through the activity and as a result of the realized learning experience). Therefore, student characteristics, including social, cultural, and economic backgrounds affect attainment by affecting individual student experiences during an activity. The factors of student background that mediate between provided "potential experiences" and student attainment are, for convenience, considered a part of the measurement strategy for achievement or the attained curriculum, since they bear a strong relationship to student attainment in most countries.

For the TIMSS, the SMSO used preliminary curriculum data to draw up the content specifications for these tests to insure concentration on those content areas that are most important to participating nations.

Three kinds of items exist in these tests: multiple choice items; items requiring short answers such as a number, phrase, diagram, or sentence; and extended response items, requiring more extended efforts to show in writing student reasoning, intermediate steps, written explanations, or other more extensive products of student performance. In addition, there are a variety of performance tasks that will be administered to a subset of the national student sample.

An analysis strategy

The examination of educational systems entails the description of their most important elements as goals and how they are implemented in classrooms and realized in student achievement. Thus, the analysis strategy must integrate the systemic components associated with the intended, implemented, and attained curriculum.

The key to this integration in the analysis strategy of the SMSO is the analytic curriculum frameworks. These frameworks form the central measurement strategy of the study. They are used to characterize curricular materials, content-specific pedagogical approaches, and student performance among other outcome measures. These important tools make the detailed interrelation of various data collections possible. This is one of the SMSO's most important contributions to systemic evaluation: the possibility of relating policy formulations of curriculum intentions to in-school implementation (school organization and classroom processes) to the attained curriculum (student achievement). The curriculum frameworks make this possible and a comprehensive set of new analytical procedures is being developed to integrate the results of the various measurements. One such integrative strategy is intended to uncover the relationship of the intended curriculum with teacher implementation and student achievement. Table 1 presents an example of an analysis table designed to illustrate data that can be used to study such relationships across countries (or subsystems).

Table 1 outlines the types of data available in TIMSS through the measurements described in the three previous sections, by which relationships between intended, implement, and attained curricula can be explored. It illustrates the exciting possibilities of the innovative design of TIMSS: we can observe patterns of national intentions as they relate to teacher presentation of content and to student achievement. Relating these patterns to other systemic characteristics and policies such as tracking, for example, provides the potential for understanding systemic change.

Exploring and describing how indicators of the intended curriculum relate to those of the implemented and attained curriculum involves the development and refinement of new indices and associated statistical procedures. This set of indicators and their links to achievement serve the goal of developing a sound system of indicators that recognizes the full complexity of the educational process.

This is the central contribution of the SMSO to the national educational policy studies—a comprehensive attempt to uncover systemic characteristics, their interrelationships with each other, and with student achievement—an enriched model for educa-

tional evaluation. The opportunity to examine a system of interrelated indicators with explicit relationships permits the examination of precisely those relationships that have been lacking in many previous methods of evaluation.

Table 1. An OTL Analysis Table

MATHEMATICAL CONTENT CATEGORY	ACHIEVEMENT DATA		TEACHER DATA			CURRICULUM AND TEXTBOOK DATA	
	no. of items	average *p* value	% of teachers reporting having covered topic in class	% of total number of mathematics class periods devoted to teaching the topic	% of teachers reporting having taught skills necessary to solve items	% of state curricul-um guides containing topic among objectives for this grade level	average % of textbooks devoted to coverage of this topic (national textbook sample)
Measurement, Perimeter, Area,							
Geometry, Congruence,							
Proportionality							
Equations and Formulas							
Data Representation and Analysis							

Note: This example illustrates some of the types of linkages that can be made in the analysis of TIMSS data, using mathematical or science content as the linking element. It shows some of the possibilities that exist to explore the relationships between curriculum policy, teachers' instructional practices, and student achievement. Such data exist for all of the almost fifty countries in the TIMSS. (This table reports only field trial data for eighth grade mathematics in the United States). The table shows that there are some content areas in which more children solve achievement test items correctly when the subject matter has been covered by their teachers and has also received considerable attention in textbooks and state curriculum guides. It also shows some content areas in which, despite teacher coverage and considerable attention in textbooks and curriculum guides, students perform poorly on related test items. Using tables such as this, in conjunction with the detailed information on teachers' instructional practices and the design of textbooks, evaluators and policy makers can uncover elements that can lead to the formulation of more effective curricular policy.

SOME ADDITIONAL EXAMPLES OF THE APPLICATION OF THE SMSO TECHNIQUES

A closer look at measuring the intended curriculum

The sources of information used by TIMSS to measure the intended curriculum are curriculum guides, textbooks, and curriculum experts. Complete descriptions of these methods can be found in a series of recent books (Schmidt and McKnight, 1995; Schmidt, McKnight, Valverde, Houang, and Wiley, 1997b; Schmidt, Raizen, Britton, Bianchi, and Wolfe, 1997c).

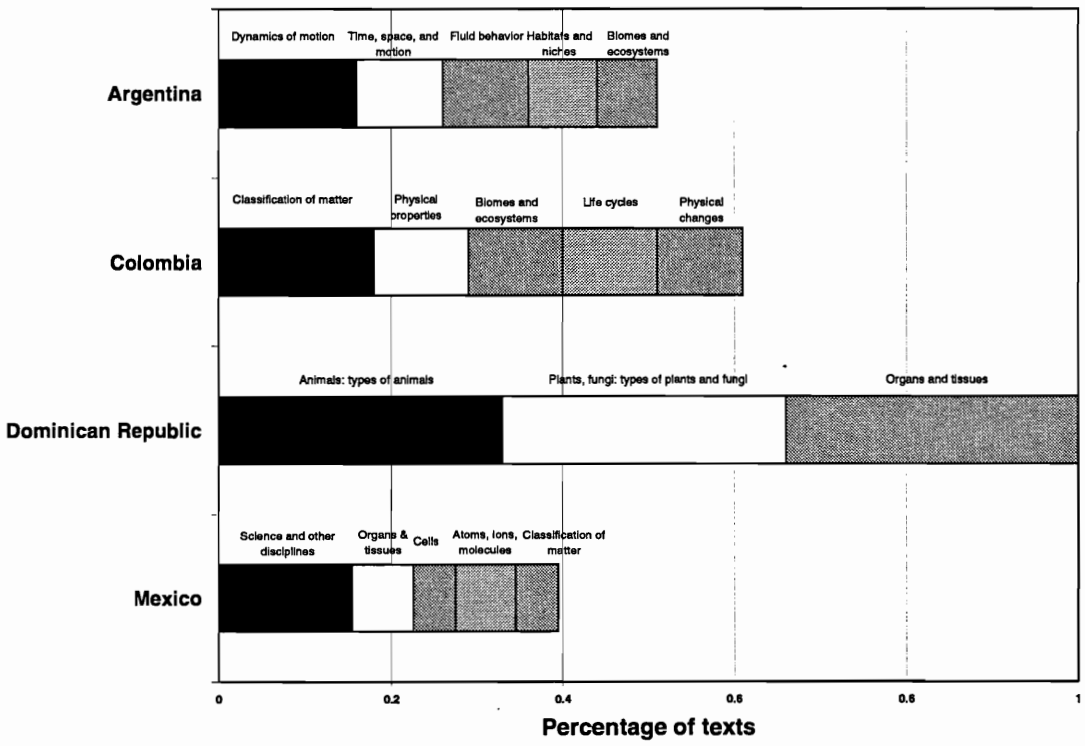
In order to understand the following examples, it is important to note that the SMSO designed a set of curriculum analysis procedures that included detailed analysis of curriculum guides and textbooks. The procedures also include tracing topic coverage across all grade levels and questionnaires for curriculum experts. These procedures are all tied to the language system of the TIMSS curriculum frameworks. For the procedure known as document analysis, a methodology of content analysis is used that involves partitioning documents into relatively homogeneous blocks, the substance of which is then coded according to the frameworks. This results in a detailed inventory of contents, performance expectations, and perspectives throughout the book. This is done for curriculum guides and textbooks at all three populations. For a small number of topics (called in-depth topics), country coders analyze the documents to provide information about the topic coverage at all grade levels, from the beginning of schooling to the end of secondary school. This task is also done for all topics in the frameworks, based on judgment rather than on formal document analysis. Experts (using documents as appropriate) provided this information. A set of mathematics and science experts from each TIMSS participant county responded to the expert questionnaire, which addresses broader issues such as reforms, patterns of governance, distribution of authority in curriculum decision making, and calculator and computer usage. This data collection has concluded, with 48 countries collecting and submitting data from more than 1,200 textbooks and curriculum guides to the SMSO.

The case of curriculum, teaching, and achievement in Colombia[3]

The Republic of Colombia participated in the recently concluded TIMSS, and the achievement results, in terms of their comparison to the other 40 countries participating in this aspect of the study, were disappointing. In science, for example, of 41 countries, 39 had mean achievement scores that were significantly higher than Colombia's. Only one country, South Africa, ranked lower. The situation for achievement in mathematics in Colombia was identical (Beaton et al., 1996a; Beaton et al., 1996b).

If the sole product of the TIMSS were these achievement scores, participation for Colombia would largely have been a waste of time. However, because the TIMSS used a measurement strategy based on a systemic model incorporating data on goals, processes, and outcomes, there are a variety of policy-relevant analyses that Colombian policy makers and researchers can pursue within the data.

Figure 4: Most Emphasized Science Topics in Grade 8 Textbooks of the Four Latin American Countries



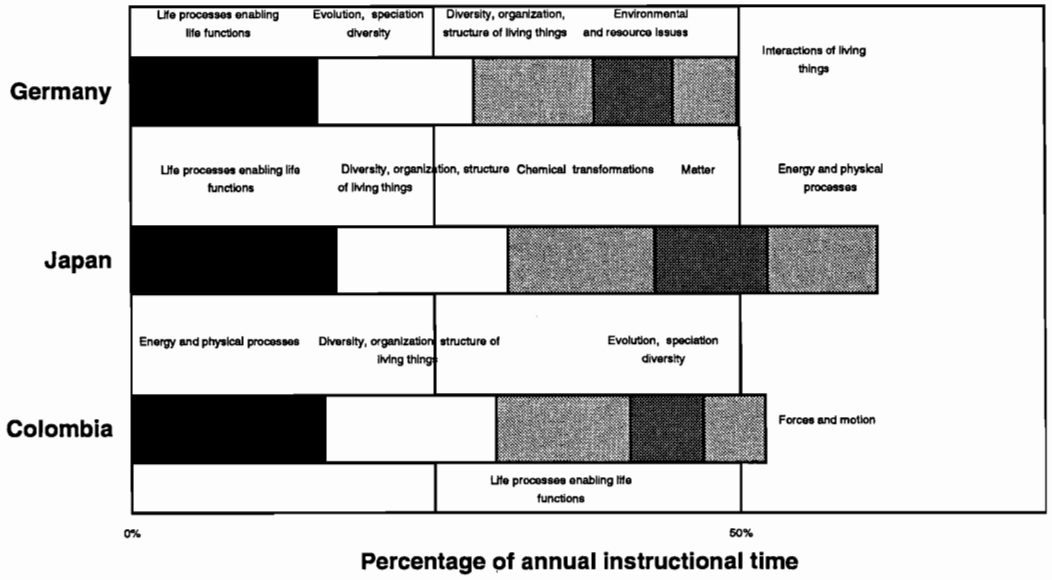
Note: Bars depict percentage of textbook space accounted for by up to five most-emphasized topics. These data are derived from a nationally-representative sample of all required or most commonly used textbooks for the full mandatory science curriculum at this grade level.

For this example, let us look first at some elements Colombia's intended curriculum as manifested in student textbooks. Figure 4 presents data on the relative topic emphases in a nationally-representative sample of Colombian eighth grade science textbooks as well as in the other three Latin American countries in the TIMSS: Argentina (province of Chaco), the Dominican Republic, and Mexico.

As can be seen, Colombian textbooks emphasize both natural and physical sciences, a situation that harmonizes with national goals made explicit in national curricular policy.

What about classroom processes? What are teachers in Colombia teaching? As Figure 5 shows, a national sample of Colombian eighth-grade teachers report emphasizing natural sciences in their instruction in close alignment with the textbooks they use. However, they do not emphasize physical sciences in a manner consistent with textbooks. Surely, these variations in opportunity to learn have consequences for student achievement. Table 2 clearly shows this.

Figure 5. Topics Most Emphasized by Teachers



Note: These are the five topics emphasized most by national samples of 8th grade teachers in Colombia, Japan, and Germany. Despite the presence of considerable material on physical science in their textbooks (see Figure 4), Colombian teachers appear to devote little instruction to these topics relative to the life sciences.

From this table it is clear that students in Colombia perform much better on those subareas of sciences that are emphasized in both textbook and classroom instruction. Further analysis in Colombia will surely be needed to prioritize a search for factors explaining the reluctance of teachers to implement the physical sciences aspects of the curriculum. A comparison of data of teaching, textbooks, and curriculum guides is likely to suggest alternatives in curriculum policy that have potential for ensuring that students are supplied with appropriate opportunities to learn the physical sciences.

The case of school-leaving examinations in the Middle East and North Africa

Since the development of the original methodologies for the TIMSS, the SMSO has had occasion to apply them to new evaluative settings, one of which was an analysis of school-leaving examinations in biology and mathematics in some countries in the Middle East and North Africa (Valverde and Schmidt, 1995).

Examinations, as stated earlier in this paper, are often intended to drive teaching and learning in educational systems. This is especially true for "gate-keeping" assessments such as those intended to provide students with a school-leaving credential or determine their passage into institutions of higher education. Of these types of tests it is particularly true that:

Table 2. Colombian Achievement Ranking in the TIMSS According to Science Topics

Contents	International Ranking
Diversity, organization, and structure of living beings	30
Life cycles	2
Biomass and ecosystems	4
Classification of matter	40
Physical properties	39
The influence of science and technology on society	6
Environmental and resource issues	17

Note: The topics in boxes are those that receive emphasis in both textbooks and teachers' instructional practices as noted in previous displays.

> "... tests exemplify educational goals... Our tests, especially in latter stages of the educational process, must reflect and form the greater society, *promoting and defining its leadership criteria as well as articulating our aspirations for it.*" (Wiley, 1982, p. 100; emphasis author's)

The World Bank, in preparing a report on education in the Middle East and North Africa (MENA), wished to investigate what school-leaving examinations evaluated using SMSO methods could reveal about the quality of education. To that end a sample of tests was collected.

The focus of the analyses performed by the SMSO was the following characteristics of the examinations: the mathematics or biology subject matter covered in each test item; and the associated expectations for student performances, or specific skills that students were intended to demonstrate in the successful solution to each item. Each of these test forms was coded using content analytic methods developed by the SMSO for the curriculum analysis component of the TIMSS. The unit of analysis that was coded was each *test task*. This term is used instead of *test item* because many individually numbered test items on each test were made up of a series of discrete tasks. That is to say, items often included more than one question to which examinees were expected to respond—each one of these questions was termed a test task.

Every task on each form for which a separate student response was intended was coded using categories from the TIMSS curriculum frameworks. The codes assigned to each task or intended response (often more than one) constitute the data from which analysis files were created. Two categories of codes were used: *content* and *performance expectations.*

Content refers simply to the mathematics or science topic areas covered by each item or task. *Performance expectation* refers to expected student performances—performances that students are expected to demonstrate while engaged in the solution of the test tasks. This latter category of codes specifies what the tests intend students to do with the mathematics or science content.

A variety of analyses were performed. Here, I would like to call attention to an analysis that has major implications for educational policy in MENA and that illustrates the policy-relevance of indicators developed by the SMSO that are not strictly related to traditional methods of evaluation.

The performance expectations—that is, what tests intended students to do with biology or mathematics—were identified for each test. The profiles that were created were then further refined to identify predominant expectations (those that represent the major emphases of the tests). The operational definition of predominance was performance expectations present in at least 70 percent of the test tasks. Table 3 presents a summary comparing mathematics school-leaving examinations with those of a European country with a long tradition in the use of such tests: the French *baccalauréate* examinations administered in Paris and Aix in 1991 and 1992.[4]

The list of MENA core expectations includes four expectations from the general areas of *knowing, using routine procedures,* and *mathematical reasoning* as they are characterized in the TIMSS frameworks. Most of these are also present in the French tests from the sample. The French forms also included performance expectations in the areas of *investigating and problem solving* and *communicating* that were not common across the MENA sample considered as a whole.

This and other analyses conducted in this study led to the discovery of a striking feature of MENA tests. When examining content and performance expectations in conjunction with each other, we found a relative lack of tasks connecting between mathematics and real-world contexts. This is remarkable because this is an area that has received considerable attention in mathematics educational reform efforts in Europe, North America, and MENA itself. In fact, tasks evaluating examinees' abilities in the area of problem solving were largely absent from tests in the MENA sample. These tests seem to indicate a concept of school mathematics as a subject largely devoted to the recognition and repetition of definitions and theorems, and the performance of algorithms and other routine procedures. There were certainly differences between the test forms and countries regarding this emphasis, but the overall trend appeared clear.

Table 3. Core Performance Expectations

MENA	France
To represent	
	Use equipment
Carry out routine procedures	Carry out routine procedures
Use more complex procedures	Use more complex procedures
	To solve
	To predict
	To verify
	To generalize
To justify and demonstrate	To justify and demonstrate
	To describe and discuss

Note: This table benchmarks core performance expectations in the mathematics school-leaving examinations of the MENA region against those in similar French examinations. Very different goals in terms of required competencies are reflected.

Stated curriculum policy in these countries clearly prioritize the promotion of critical thinking, practical problem solving, and information processing skills, as is the case of the Kingdom of Jordan (Billeh, 1996). Yet these high-stakes, school-leaving examinations do not prioritize such skills, making it extremely unlikely that they will receive high priority with teachers and students that are judged against them.

Uncovering these types of inconsistences between two aspects of the intended curriculum—tests and curriculum policy—has proven of use to both individual nations and the World Bank as they evaluate policy options for reforming education in ways that will propitiate the development of competitive economies.

SOME CONCLUSIONS: EVALUATION AND EFFECTIVE CURRICULUM POLICY MAKING

Several implicit causal assumptions underlie public and political interest in education throughout the Americas: they hold that improved educational delivery systems will result in greater achievement for a greater number of students. This, in turn, is believed to result in a greater number of better-prepared students entering the labor force, resulting in a more competitive economic performance in the global economy. Increased general knowledge of school subjects is understood to stimulate innovations across all sectors of society and the economy, resulting in a more

informed electorate and consumer population, increased productivity in the general workforce, and enhanced competitiveness in national industry. This view has been stated forcibly and repeatedly—that education is one of the single most important human capital investments a country must make (Álvarez, 1997; Becker, 1964; Board on International Comparative Studies in Education, 1993; Ministerio de Educación Nacional de Colombia, 1990; Ministerio de Educación Pública de Costa Rica, 1996; The World Bank, 1995).

Beyond this argument linking personal educational development through schooling to national economic growth and competitiveness, the educational policy debate has urged us to consider educational institutions as a system.

This system is seen as amenable to reform through informed policy making. Sets of evaluation priorities have been established by a variety of politicians in different nations—as well as by international organizations such as the OECD and UNESCO. Approaches to educational evaluation that include the educational process in their consideration of educational goals and outcomes can make important contributions to educational policy making. This is possible by providing indicators that will enable us to understand the links between national goals, what occurs in schools and classrooms, and student achievement. The policy relevance of an approach, such as that of the SMSO, is its focus on the notion of educational opportunity as an important outcome of educational systems, and its offering a model of educational opportunity that permits its description in systemic terms.

In the preceding pages I have described and illustrated some of the benefits of the curriculum-sensitive evaluation strategy developed by the SMSO. I have also explored how indicators of process can provide information relevant for policy making. Some of the most important benefits offered by the indicator system described here, relate to increasing the capacity to formulate effective curriculum policy.

Curriculum policy is currently undergoing substantive revisions in many countries. Perhaps one of the most important problems that policy makers must struggle with is that of reconciling the diverse messages concerning curriculum that are sent by its most important instruments: curriculum standards or frameworks, programs of study, textbooks, and achievement assessments.

The system of measures described here permits the evaluator to characterize, in considerable detail, how each of these policy instruments conveys messages concerning the curriculum to each person and agency in the system. It can reveal, for example, whether the objectives set forth in curriculum frameworks are supported by achievement tests and textbooks, providing an important tool for those systems in which some elements of curriculum policy are entirely the responsibility of central authorities (e.g., curriculum frameworks) and other elements are left to other people (e.g., textbooks developed by private publishing houses). They therefore provide a tool either for the elaboration of coherent policies regarding textbook development (for those systems in which textbooks are developed by central authorities), adoption (for those systems that produce a list of

"authorized" commercially-developed textbooks), or provide useful terms of reference where countries purchase textbooks within a program of competitive bidding.

Curriculum frameworks, while setting important educational goals for the system, are typically not designed to suggest specific implementation strategies. Such implementation strategies (that is, suggestions for the concrete actions to be taken by specific people or organizations to realize pedagogical goals) are contained in other policy instruments such as programs of study and textbooks. Designing more effective programs of study and textbooks is also enhanced by evaluation systems such as the one designed by the SMSO. Since the evaluation of educational processes provides information on instructional practices and how they affect the achievement of different pedagogical objectives, they provide a good basis for the design of such instruments. The inclusion of such measures in national evaluation programs provides a vital feedback mechanism that will be of use to curriculum developers, textbook writers, test developers, and others in the system in the modification and perfection of programs of study, textbooks, and tests over a period of time.

These indicators are also of obvious use in the design of effective policies regarding teacher preservice and inservice training. A remarkable feature of current curriculum policy in much of the Americas is that despite a call for a considerable change in pedagogy, there is little explicit consideration of how the educational system will produce the teachers required to deliver the more demanding curricula. In uncovering effective and ineffective implementation strategies—as they correspond to specific pedagogical goals relating to expected student performances and/or the content of school subjects—such indicators can provide important information that can influence teacher preparation curricula and the design of inservice programs targeted at addressing shortcomings in the delivery of educational opportunities.

Traditional evaluation systems, by concentrating the attention of policy makers and the public on the "horse races" or "achievement Olympics" that compare the test scores of children, or the mean scores of classrooms, schools, provinces, and even nations, reify such rankings and confuse the important issues. Awareness of poor performance may focus intention on improvement.

However, by themselves, such rankings are purposeless, confusing, and may in fact be destructively alarmist. They may lead at best to simplistic proposed solutions and at worst, to ill-conceived policies with deleterious consequences. What policy makers require are data on how a variety of educational goals are pursued by elements of the educational system, and on which systemic characteristics, instructional practices, and other educational processes serve best to further these different goals. This type of information will make a positive contribution to the public debate, helping to direct our attention to consider policies with the potential to promote our goals.

This is evaluation as conceived in modern policy analysis: "expanding the task of evaluation beyond the mere measurement of outcomes to their causes" (Pressman and Wildavsky, 1984: p. xv). Viewed in this way, such approaches to educational evaluation

can provide a rich source of information to inform key national policy issues of the late 1990s and beyond throughout the Americas.

NOTES

[1] The SMSO was a grant awarded to Michigan State University as the U.S. Research Center for the TIMSS, which coordinated research and development work with the national research centers of a subgroup of nations participating in the TIMSS: Japan, Norway, Switzerland, Spain, France, and the United States. The purpose of the SMSO was to develop the following for the TIMSS: a set of curriculum frameworks, a conceptual model to guide instrument development and data analysis, questionnaires for teachers and students, techniques for measuring and studying intended curricula, and the specifications (blueprints) for the TIMSS achievement tests.

[2] Detailed reports on the instrumentation are provided in our book (Schmidt et al., 1996), and in a series of technical reports from the SMSO Center at Michigan State University. For further information please contact the author at: SMSO, 457 Erickson Hall, East Lansing, Michigan 48824-1034.

[3] I wish to acknowledge my debt to my colleague, Dr. Carlos Jairo Díaz, for his permission to use Colombian TIMSS data for these analyses. All questions regarding the specific circumstances of Colombian participation in the TIMSS should be directed to Dr. Díaz at: Multitaller de Materiales Didácticos, Universidad del Valle, Ciudad Universitaria Meléndez, Apartado Aéreo 25360, Cali, Colombia.

[4] Data for these French examinations were provided by National Center for Improving Science Education, Washington D.C. The French data were coded by Dr. John Dossey, Illinois State University. Additional information concerning these examinations can be found in: Britton, Edward D., and S. A. Raizen. *Examining the Examinations: An International Comparison of Science and Mathematics Examinations for College-Bound Students.* Boston/Dordrecht/London: Kluwer Academic Press, 1996.

REFERENCES

Airasian, P.W., and Madaus, G.F. (1983). "Linking Testing and Instruction: Policy Issues." *Journal of Educational Measurement,* 20(2), 103-118.

Álvarez, B. (1997). "Naturaleza y Contexto de las Reformas Educativas de Final de Siglo." In B. Álvarez and M. Ruiz-Casares (eds). *Senderos de Cambio: Génesis y Ejecución de las Reformas Educativas en América Latina y el Caribe* (Vol. 1, pp. 1-22). Washington DC: Advancing Basic Education and Literacy Project, Academy for Educational Development.

Álvarez, B., and Ruiz-Casares, M. (eds). (1997). *Senderos de Cambio: Génesis y Ejecución de las Reformas Educativas en América Latina y el Caribe.* Washington DC: Advancing Basic Education and Literacy Project, Academy for Educational Development..

Apple, M.W. (1990). *Ideology and Curriculum.* (Second ed.). New York: Routledge.

Apple, M.W. (1992). "Do the Standards Go Far Enough? Power, Policy, and Practice in Mathematics Education." *Journal for Research in Mathematics Education, 23*(5), 412-431.

Beaton, A.E., Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., Smith, T.A., and Kelly, D.L. (1996a). *Mathematics Achievement in the Middle School Years: IEA's International Mathematics and Science Study.* Chestnut Hill, MA: TIMSS International Study Center.

Beaton, A.E., Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., Smith, T.A., and Kelly, D.L. (1996b). *Science Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study.* Chestnut Hill, MA: TIMSS International Study Center.

Becker, G.S. (1964). *Human Capital: A Theoretical and Empirical Analysis.* New York: Columbia University Press.

Billeh, V.Y. (1996). *Educational Reform in Jordan: An Analytical Overview.* Amman, Jordan: National Center for Human Resources Development.

Board on International Comparative Studies in Education. (1993). *A Collaborative Agenda for Improving International Comparative Studies of Education.* Washington, DC: National Academy Press.

Burstein, L. (1993). "Studying Learning, Growth, and Instruction Cross-Nationally: Lessons About Why and Why Not to Engage in Cross-National Studies." In L. Burstein (ed). *The IEA Study of Mathematics III: Student Growth and Classroom Processes.* New York: Pergamon Press.

Burstein, L., Aschbacher, P., Chen, Z., and Sen, Q. (1990). *Establishing the Content Validity of Tests Designed To Serve Multiple Purposes: Bridging Secondary-Postsecondary Mathematics.* Los Angeles: University of California Los Angeles, Center for Research on Evaluation, Standards, and Student Testing.

Carlsen, W.S. (1991). "The Construction of Subject Matter Knowledge in Primary Science Teaching." In J. Brophy (ed). *Teachers' Subject Matter Knowledge and Classroom Instruction* (Vol. 2). Greenwich, CT: JAI Press.

Clune, W.H. (1993). "The Best Path to Systemic Educational Policy." *Educational Evaluation and Policy Analysis, 15*(3), 233-254.

Cohen, D.K. (1988). "Teaching Practices: Plus que ça Change ..." In P. Jackson (ed). *Contributing to Educational Change: Perspectives on Research and Practice* (pp. 27-84). Berkeley, CA: McCutchan.

Elmore, R.F., and Fuhrman, S.H. (1994). "Governing Curriculum: Changing Patterns in Policy, Politics, and Practice." In R.F. Elmore and S.H. Fuhrman (eds). *The Governance of Curriculum* (pp. 1-10). Alexandria, VA: Association for Supervision and Curriculum Development.

Guiton, G., and Oakes, J. (1995). "Opportunity to Learn and Conceptions of Educational Equality." *Educational Evaluation and Policy Analysis, 17*(3), 323-336.

Linn, R.L. (1987). *State-by-State Comparisons of Student Achievement: The Definition of the Content Domain for Assessment* (275). Los Angeles: University of California Los Angeles, Center for Research on Evaluation, Standards, and Student Testing.

McDonnell, L. M. (1995). Opportunity to Learn as a Research Concept and a Policy Instrument. *Educational Evaluation and Policy Analysis, 17*(3), 305-322.

Ministerio de Educación Nacional de Colombia. (1990). *Matemáticas: Marco General y Propuesta de Programa Curricular–Octavo Grado de Educación Básica.* Bogotá, Colombia: Ministerio de Educación Nacional.

Ministerio de Educación Pública de Costa Rica. (1996). *Programa de Estudios: Matemática: Educación Diversificada.* San José, Costa Rica: Ministerio de Educación Pública.

Muthén, B., Huang, L.-C., Jo, B., Khoo, S.-T., Goff, G.N., Novak, J.R., and Shih, J.C. (1995). "Opportunity-to-Learn Effects on Achievement: Analytical Aspects." *Educational Evaluation and Policy Analysis, 17*(3), 371-403.

Noble, A.J., and Smith, M.L. (1994). "Old and New Beliefs About Measurement-Driven Reform: Build It and They Will Come." *Educational Policy, 8*(2), 111-136.

O'Day, J.A., and Smith, M.S. (1993). "Systemic Educational Reform and Educational Opportunity." In S.H. Fuhrman (ed). *Designing Coherent Educational Poliy* (pp. 250-312). San Francisco: Jossey-Bass.

Porter, A.C. (1993, June-July). "School Delivery Standards." *Educational Researcher, 22,* 24-30.

Porter, A.C., Archbald, D.A., and Tyree, A.K., Jr. (1991). "Reforming the Curriculum: Will Empowerment Policies Replace Control?" In S.H. Fuhrman and B. Malen (eds). *The Politics of Curriculum and Testing* (pp. 11-36). London: The Falmer Press.

Pressman, J.L., and Wildavsky, A. (1984). *Implementation.* (Third ed). Berkeley: University of California Press.

Robitaille, D.F., Schmidt, W.H., Raizen, S., McKnight, C., Britton, E., and Nicol, C. (1993). *Curriculum Frameworks for Mathematics and Science.* (Vol. 1). Vancouver, Canada: Pacific Educational Press.

Schmidt, W.H., Jorde, D., Barrier, E., Gonzalo, I., Moser, U., Shimizu, K., Sawada, T., Valverde, G.A., McKnight, C., Prawat, R.S., Wiley, D.E., Raizen, S.A., Britton, E.D., and Wolfe, R.G. (1996). *Characterizing Pedagogical Flow: An Investigation of Mathematics and Science Teaching in Six Countries.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

Schmidt, W.H., and McKnight, C.C. (1995). "Surveying Educational Opportunity in Mathematics and Science: An International Perspective." *Educational Evaluation and Policy Analysis, 17*(3), 337-353.

Schmidt, W.H., McKnight, C.C., Raizen, S.A., Jakwerth, P.M., Valverde, G.A., Wolfe, R.G., Britton, E.D., Bianchi, L.J., and Houang, R.T. (1997a). *A Splintered Vision: An Investigation of U.S. Science and Mathematics Education.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

Schmidt, W.H., McKnight, C.C., Valverde, G.A., Houang, R.T., and Wiley, D.E. (1997b). *Many Visions, Many Aims: A Cross-National Investigation of Curricular Intentions in School Mathematics.* (Vol. 1). Dordrecht, The Netherlands: Kluwer Academic Pulblishers.

Schmidt, W.H., Raizen, S.A., Britton, E.D., Bianchi, L.J., and Wolfe, R.G. (1997c). *Many Visions, Many Aims: A Cross-National Investigation of Curricular Intentions in School Science.* (Vol. 2). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Secretaría de Educación Pública de México. (1993). *Educación Básica Secundaria: Plan y Programa de Estudios.* México D. F.: Secretaría de Educación Pública.

Shavelson, R.J., and Webb, N.M. (1995). "On Getting It Right." *Educational Evaluation and Policy Analysis, 17*(3), 275-279.

Slavin, R.L. (1994). "Effective Classroom, Effective Schools: A Research Base for Reform in Latin American Education." In J.M. Puryear and J.J. Brunner (eds). *Education, Equity and Economic Competitiveness in the Americas* (pp. 7-28). Washington DC: Organization of American States.

Smith, M.S., Fuhrman, S.H., and O'Day, J. (1994). "National Curriculum Standards: Are They Desirable and Feasible?" In R. F. Elmore and S. H. Fuhrman (eds). *The Governance of Curriculum* (pp. 13-29). Alexandria, VA: Association for Supervision and Curriculum Development.

Survey of Mathematics and Science Opportunities. (1992a). *TIMSS Mathematics Curriculum Framework* (38). East Lansing, MI: SMSO.

Survey of Mathematics and Science Opportunities. (1992b). *TIMSS Science Curriculum Framework* (37). East Lansing, MI: SMSO.

The World Bank. (1995). *Claiming the Future: Choosing Prosperity in the Middle East and North Africa.* Washington DC: The World Bank.

Valverde, G.A., and Schmidt, W.H. (1995). *An Exploratory Analysis of the Content and Expectations for Student Performance in Selected Mathematics and Biology School-Leaving Examinations from the Middle East and North Africa* . Washington DC: The World Bank.

Wiley, D.E. (1982). "The Vicious and the Virtuous: ETS and College Admissions." *Contemporary Education Review,* 1(2), 85-101.

Zedillo Ponce de León, E. (1993). "Presentación." In Ministerio de Educación Pública (ed). *Educación Básica Secundaria: Plan y Programa de Estudios* (pp. 7). México DF: Ministerio de Educación Pública.

CHAPTER 4

INTERNATIONAL MONITORING OF THE GOALS OF HUMAN DEVELOPMENT: THE CASE OF THE SECRECTARIAT PRO TEMPORE OF THE AMERICAS

María Inés Cuadros Ferré

The debate over educational standards has transcended national borders and is fast becoming a topic of interest in the field of international politics. Few experiences exist, however, related to monitoring educational goals in various countries. The Secretariat Pro Tempore is one such experience from which useful lessons can be drawn. This paper summarizes the principal agreements and goals in the field of children's issues subscribed to by governments of countries of the Americas following approval of the International Convention on the Rights of the Child. It describes the specific mechanism created by the governments in 1994, now known as the Secretariat Pro Tempore, for monitoring the agreements at the regional level. In addition, it reviews the various activities carried out during its initial implementation period (1994-1996), and analyzes its achievements and difficulties. Finally, it presents several recommendations and proposals aimed at encouraging the formulation of commitments and the establishment of international educational goals, while offering reflections on the future of processes for monitoring mechanisms of this type.

BACKGROUND

The International Convention on the Rights of the Child and the World Summit for Children

On November 20, 1989, the United Nations approved the International Convention on the Rights of the Child, which went into effect on September 2, 1990. The intent of the Convention was to transform the child from an object of special protection into the subject of a broad range of rights and freedoms (for the text of the Convention as it affects the area of education, see Table 1). In addition, the Convention recognizes the dignity of the child as a person, and consequently compliance with the rights of the child becomes obligatory and legally binding. As of early 1997, the Convention had been ratified by 190 nations.[1]

To promote the Convention, the World Summit for Children was held in New York, under the auspices of the United Nations, on September 30, 1990. During that summit, 71 heads of state and government and representatives of 88 countries signed the World Declaration on the Survival, Protection, and Development of Children. By doing so, they committed themselves to promoting the ratification of the International Convention on the Rights of the Child (thus making it a legally-binding instrument in their respective countries) and to making every possible effort to create, by the year 2000, improved living and developmental conditions for children. The Declaration established the commitment to implement a program based on ten core points, together with their corresponding goals, to be orchestrated through a World Plan of Action. This Plan defines sectoral support objectives, grouped by subject matter into the areas of women's health and education, nutrition, child health, water and sanitation, basic education, and protection for children and adolescents living in extremely difficult conditions.

With regard to education, the Plan defined the following as two of the most important goals to be achieved by the year 2000: to guarantee "universal access to basic education and completion of primary school for at least 80 percent of school-age children, and reduction of the illiteracy rate among adults to less than half the level recorded in 1990, with special emphasis to be given to literacy training for women."

To achieve these goals, four specific objectives were established:

- Expansion of early childhood development activities, including appropriate low-cost family- and community-based interventions;
- Universal access to basic education and completion of primary schooling by at least 80 percent of school-age children by means of either school-based or nonacademic education with a comparable level of learning, with an emphasis on reducing current existing disparities between girls' and boys' education;
- Reduction in the rate of adult illiteracy to at least 50 percent of the level recorded in 1990, with emphasis on literacy training for women; and
- Increased accumulation by individuals and families of the knowledge, techniques, and values necessary to lead a better life, which are to be provided to them through all available educational channels. These channels include mass media and other modern and traditional forms of social action and communication, the effectiveness of which would be measured as a function of the behavioral changes recorded.

Table 1. International Convention on the Rights of the Child: Rights in the Field of Education

Article 28
1. State Parties recognize the right of the child to education, and with a view to achieving this right progressively and on the basis of equal opportunity, they shall, in particular: (a) make primary education compulsory and available free to all; (b) encourage the development of different forms of secondary education, including general and vocational education, make them available to every child, and take appropriate measures such as the introduction of free education and offering financial assistance in case of need; (c) make higher education accessible to all on the basis of capacity by every appropriate means; (d) make educational and vocational information and guidance available and accessible to all children; (e) take measures to encourage regular attendance at schools and the reduction of drop-out rates.
2. State Parties shall take all appropriate measures to ensure that school discipline is administered in a manner consistent with the child's human dignity and in conformity with the present Convention.
3. State Parties shall promote and encourage international cooperation in matters relating to education, in particular with a view to contributing to the elimination of ignorance and illiteracy throughout the world and facilitating access to scientific and technical knowledge and modern teaching methods. In this regard, particular count shall be taken of the needs of developing countries.

Article 29
1. State Parties agree that the education of the child shall be directed to: (a) the development of the child's personality, talents, and physical abilities to their fullest potential; (b) the development of respect for human rights and fundamental freedoms, and for the principles enshrined in the Charter of the United Nations; (c) the development of respect for the child's parents, his or her own cultural identity, language and values, for the values of the country in which the child is living, the country from which he or she may originate, and for civilizations different from his or her own; (d) the preparation of the child for responsible life in a free society, in the spirit of understanding, peace, tolerance, equality of sexes, and friendship among all peoples, ethnic, national, and religious groups, and persons of indigenous origin; (e) the development of respect for the natural environment.

The Convention and the National Plans of Action for Children in America

In the Americas, the Declaration of Heads of State produced a very significant impact that led to the early ratification of the Convention (to date, 33 of the 34 nations have given their ratification) and to the formulation of National Plans of Action (NPAs) for Children.[2] With regard to the field of development and education, the NPAs include the following subject areas: (a) early stimulation programs; (b) initial education programs; (c) intercultural school projects; (d) strengthening of the coverage, access, and quality of basic education; (e) development of ten-year education plans; (f) educational reform projects; and (g) development of monitoring systems.

In order to monitor and analyze the progress achieved in complying with the obligations undertaken by the State Parties to the Convention, the Committee on the Rights

of the Child was created. Countries are required to submit to the Committee reports on measures taken to implement the rights recognized in the Convention and on progress achieved with regard to the enjoyment of such rights.

All country reports submitted to the Committee include a chapter on progress achieved and difficulties encountered in the field of education. In addition, most of the reports submitted by nongovernmental organizations contain critical analyses of the degree of compliance with the right to education. In this regard, their concern is focused on the deterioration of public schools in a number of countries, the deficiencies observed in terms of the quality of education, the high rates of school dropout and grade repetition, the scant attention paid to the initial education of extremely poor children, the very marked differences in terms of access to education in urban areas as opposed to rural areas, the lack of equity between the sexes, the lack of development of pedagogical projects where such projects actually exist, the mistreatment to which boys and girls are subjected in school, the association between working children and children who drop out of school, and the difficulties experienced by bilingual children with regard to access to education in their native languages as a result of the failings of educational systems.

In 1992 the nations of the Americas, with support provided from agencies of the United Nations, initiated a series of periodic meetings designed to promote the formulation of National Plans of Action for Children and to monitor the achievements recorded, and difficulties encountered, in their application. The first such meeting took place in Brazil: in April 1992, with support made available by the Pan American Health Organization (PAHO/WHO), an invitation was extended to representatives of the health sector and parliamentarians from countries of the Americas to formulate or revise the National Plans of Action for their countries. Subsequently, following the initiative of the Government of Mexico, a Latin American Ministerial Meeting on Children and Social Policy was held to assess progress toward compliance with the goals and analyze such issues as intersectoral coordination, intercountry cooperation, and the financing of the Plans. At the conclusion of the meeting, the attendees signed the Declaration of Tlatelolco, which reaffirmed the unpostponable commitment of the governments to comply with the initiatives outlined at the World Children's Summit.

As a result, in April 1994 the Second American Meeting on Children and Social Policy was convened in Bogotá, Colombia. Ministers and representatives from twenty-eight countries of the Americas responsible for the agreements in support of children in the Region attended the meeting. The principal topics addressed involved the institutionalization of social policy and the National Plans of Action for Children, the status of goals for children at the midpoint of the decade and the corresponding challenges, decentralization and municipalization, financing, and international and horizontal cooperation. A significant milestone occurred with the signing of the Nariño Pact. This agreement was instrumental in launching what came to be known as the Secretariat and preceded the drafting of the Santiago Agreement, which was signed in August 1996 in Santiago de Chile on the occasion of the Third American Meeting on Children and Social Policy and is currently in force.

The Nariño Pact

The Nariño Pact reaffirms the obligation to comply with agreements signed at the World Summit, declares that the crux of social and economic policy must be human development, identifies the need to invest in children in order to ensure sustainable and equitable human development, and proposes measures to be taken by governments to overcome the structural factors of poverty and propitiate the efficiency and productivity of the economy. The Pact is structured around six sections, with the first referring to the goals of human development; the second to institutionalization; the third to commitments in the field of decentralization, public management, and social participation; the fourth to the objectives of international financing and cooperation; and the fifth to regional monitoring mechanisms. The sixth section presents a number of final considerations, while the appendix spells out the way in which the regional monitoring mechanism is to be implemented.

In the field of education, the agreement identifies clearly-defined paths to be followed by the countries of the Americas. With regard to goals, in addition to confirming those established at the World Summit, the agreement stresses the need to develop an initial community and family-based education, with a commitment to identifying methodologies and strategies for expanding coverage and developing alternative care models for the youngest age group of children. In addition, it calls attention to the quality of education. The Pact commits to enhancing the quality of primary education by introducing curriculum reforms; increasing investments in infrastructure, appropriate texts, and teacher training; and ensuring that the school period is adequate to address countries' needs and ensure effective learning. Lastly, it identifies technical-vocational education and training for youths as a necessary goal for the countries of the Americas.

Beyond the goals in the field of education *per se*, the above-mentioned agreements are vitally important for the process of educational development. Thus, for example, the signatories resolve to support decentralization on the regional and municipal levels. They reaffirm the extreme importance assigned to promoting the participation of a range of social groups implementing the National Plans of Action, which include educational plans. Additionally, the agreements call for continued efforts to increase social investments at the country level and to improve efficiency, effectiveness, and equity. In this way, they contribute to guaranteeing the rights and basic needs of the hemisphere's poorest of the poor.

THE AMERICAN SECRETARIAT PRO TEMPORE, 1994-1996[3]

In creating the monitoring mechanism, the Pact launched what is today the Secretariat Pro Tempore for monitoring agreements in support of children in the Americas. The primary functions of the Secretariat include promoting the exchange of successful child-related social development experiences; promoting the intraregional exchange of information on methodologies and instruments of social policy; disseminating on a broad scale information related to social issues linked particularly to children; and supporting the creation or strengthening of national or subregional information systems.

The government of the host country for each American Meeting on Children and Social Policy is responsible for coordinating the established regional mechanism, up until the next convened meeting. Under this agreement, the Government of Colombia was responsible for launching the American Secretariat Pro Tempore, which was installed in the National Planning Department and placed under the direction of that office, which in addition was responsible for coordinating the National Plan of Action for Children. Its initial responsibilities consisted of defining both its role and activities, as well as the strategy and mechanism that would govern its relationship with the various countries, cooperation agencies, and nongovernmental organizations.

The Secretariat was defined as a mechanism of support established by American countries for the purposes and actions previously identified in the Nariño Pact. Due to its nature as a support mechanism for monitoring the Plans of Action for Children, which have their origin and basis in the International Convention on the Rights of the Child and in the World Declaration on the Survival, Protection, and Development of Children, the Secretariat deemed it necessary to carry out all of its activities from the perspective of those mandates.

Additionally, because UNICEF was acting as Secretariat of the Interagency Committee,[4] it was felt that work would be conducted directly through this entity. The Secretariat Pro Tempore would work through UNICEF to establish relationships with international agencies and organizations.

Monitoring of goals

It was believed that each goal identified in the Pact should be expressed by means of one or more indicators, so that the countries, with support from the Secretariat and using either quantitative or qualitative parameters, would be able to monitor countries' commitment in a meaningful way. Toward this end, a table of indicators was developed, together with the corresponding definitions and formulas for obtaining them. Subsequently, the countries were asked to administer a questionnaire with information related to the goals, recommended indicators, most recent available information, source from which such information was to be obtained, frequency of data, level of disaggregation possible, and value of the goal by country for the years 1995 and 2000.

Twenty-four countries of the hemisphere administered the survey. The analysis conducted by the Secretariat made it possible to identify the achievements and difficulties experienced by the countries as a whole with regard to the information systems necessary for supervising the goals related to children. The survey results revealed that those indicators of educational evolution referred to as traditional—such as rates of schooling, dropouts and number of teachers—were in effect available in the region. However, new indicators related to the quality of education and system results were not included in the information systems. There were no indicators to measure, for example, acquisition of skills, number of hours of learning, or investments in text books. Without such indicators, it is simply not possible to monitor the results of the education reform programs in which most countries find themselves currently immersed.

As of the date of the survey (first half of 1995), the sources of data for educational goals were primarily the National Education Ministries or Educational Planning Offices. Additional sources included national statistical departments or institutes, the educational information system, and UNESCO. The principal instrument used was the administrative registry, together with censuses and both ongoing and specific time-limited surveys.

Although subnational disaggregation was fairly widespread in the region, disaggregation by gender and age was at an incipient level. It was believed that this deficiency could be attributed to the inappropriateness of the instruments employed, to the poor processing of questionnaires, or to the fact that appropriate use was not being made of all data available at the time the questionnaire was administered. Whatever the case, considerable attention must be given to this finding, since generational analyses are much more illustrative for sector planning purposes and since equity concerns dictate the need for gender-based analysis.

The general conclusions revealed that (1) health and education indicators were more readily available than civil rights indicators; (2) in all areas, including those indicators considered to be traditional, standardization was absent, thus preventing comparisons among countries; (3) the greatest lack of standardization concerned relevant age ranges, rate increase factors, and the categories included; (4) in general, the various indicators lacked corresponding goals, regardless of whether the countries had ratified the international agreements or not; and (5) there is no clear idea of what an information system for monitoring goals actually is or how it should operate.

Salient to the general recommendations was the need to create an institutional awareness of the significance of goal monitoring and of its usefulness not only for individual countries but for all countries as a whole. In addition, these priorities were assigned: continuing the effort to design and standardize indicators, strengthening institutional abilities in information processing in order to better utilize the potential of the various sources, and supporting countries in establishing appropriate information systems.

The Secretariat Pro Tempore, together with the Government of Guatemala and UNICEF, prepared and held a technical meeting in August 1995. The meeting was intended to analyze the mechanisms for monitoring the goals of the National Plans of Action with regard to sources of information, relevance of the proposed indicators, and frequency and regularity with which progress and achievements were to be assessed, as well as to promote the exchange of successful experiences with information systems among the countries of the region.

The presentations made during this meeting provided evidence of how the household surveys and the multiple indicators survey constitute valuable instruments; identified some of the challenges that can occur when measuring inputs, processes, and results in the field of education; provided information on the development, use, and presentation of databases; and stressed the importance of using educational information for policy decision making.

Participants in turn identified the need to generate strict and imaginative educational indicators and to create indicators that would measure the impact of educational achievements on the decrease in poverty. They also pointed to existing difficulties regarding the establishment of coverage and methodologies in preschool education, due especially to the private nature of the latter in most countries, and the challenge of determining the educational status of those children not associated with the formal system.

Among the recommendations developed at the meeting, the following stand out:

- strengthen information systems so that they will be able to provide the inputs necessary for conducting periodic assessments of progress toward achievement of the goals outlined in the programs (assessment results would facilitate decision making aimed at generating actions and orienting social expenditures in accordance with the actual needs of the population);
- improve and normalize the availability, reliability, timeliness, and use of information related to the status of children;
- strengthen social information systems, enhance production, analysis, and utilization at the national, regional and local levels, and conduct an analysis of the social situation by sex, ethnic group, and at-risk population group;
- emphasize participation of civil society in the production, analysis, and use of the information; and
- collect and disseminate information for improving the ability of children and youth to participate, freely express themselves, and give and receive information.

PROGRESS TOWARD THE ACHIEVEMENT OF THE GOALS OF HUMAN DEVELOPMENT

At the conclusion of the term during which it was under the responsibility of the Government of Colombia, the Secretariat Pro Tempore analyzed the activities carried out to date by the countries of the Americas and measured their progress toward compliance with the goals. To prepare its report, the Secretariat reviewed the following information: (a) responses provided by the countries to an evaluation questionnaire designed by the Secretariat; (b) the mid-decade progress reports submitted by countries to UNICEF; (c) government and nongovernmental reports submitted by countries to the Committee on the Rights of the Child; and (d) specialized reports available through the agencies of the United Nations, as well as documents pertaining to the Secretariat itself.

Growth and development

The country reports identified two modes of care for boys and girls between the ages of zero and five:

Monitoring programs. These are programs normally conducted by the health sector in which growth and development monitoring activities are carried out. Only two coun-

tries[5] had in place a validated scale for systematically measuring child growth and development.

Day-care programs. At least three modes were identified in such programs, namely community homes and day-care homes, which are characterized by a high degree of community participation; care provision networks based on the coordination between government efforts at the international level on the one hand and local independent initiatives and the initiatives of nongovernmental organizations on the other (these two modes include nutrition and protection components and normally receive some type of government subsidy); and preschool child-care programs, designed primarily for the middle and upper classes and organized by the private sector.

The demand for child-care continues to grow because most women work and require institutions or organizations to assist them by providing care and nurture for their children. The countries of the region identified as a priority the training and qualification of both parents and child-care providers (those who care for the youngest age group). Making training a priority would support child development beginning in the first months of life and demonstrate efforts to restore the educational and protectional function of family and community.

Initial and preschool education and basic preparation

There is a tendency in the region toward the strengthening of initial education programs, as evidenced in the expansion and diversification of programs and methodologies based on nonformal schemes characterized by a high degree of community participation. The countries stress the benefits of supporting such programs, as evidenced in their impact on child development, greater facility for accessing the formal education system, and reduction in the rates of grade repetition. Nevertheless, only 20 percent of children between the ages of five and six have access to such programs.[6]

Basic education

The countries of Latin America and the Caribbean simultaneously record both high levels of primary enrollment and high rates of grade repetition and dropout. Thus, net access to basic primary education is 86 percent for children between the ages of six and eleven, with the lowest level (75 percent) found in the countries of Central America and the highest (93 percent) in the English-speaking Caribbean. Forty-two percent of those who enroll in school repeat first grade, while 30 percent repeat second grade. The average rate of grade repetition for all primary school grades is 30 percent, with costs totaling US$3.5 billion.

The principal causes associated with the fact that twenty-two million children in the region repeat school grades include the poor quality of the educational system, the scant availability of resources in the schools, the absence of teaching aids, and the lack of teacher training. Additional causes include the characteristics of the children themselves in terms of their physical development, nutrition, adaptation, and psychological

development. A significant portion of students who remain in the system achieve significantly below those in other areas of the developing world. The highest levels of achievement have been observed in the English-speaking region of the Caribbean and among private school students in other countries of the area.

Enhancing the quality of education is one of the challenges facing the region and will involve, among other things, the following: stressing the learning of basic skills in reading, writing, and arithmetic; promoting the development of materials and guides that will respond to the needs of existing and diverse student groups; offering more appropriate and higher quality education in rural areas and to existing ethnic groups; reinforcing the management capability of territorial entities; and guaranteeing the provision of training, improved pay, and incentives for teachers.

Secondary education

It is in this stage that the effects of school dropout and lack of access to the system by the large majority of adolescents become increasingly more evident. Although availability of secondary education increased from 14 percent in 1960 to 55 percent in 1993, almost half of the continent's adolescent population does not have access to this level of schooling.

The implications of the low rates of coverage are varied and significantly affect the countries' level of development. On the one hand, economic processes demand increasingly qualified human capital, while on the other the low education levels lead to certain negative effects that affect future generations. For example, families are started at an earlier age, and have more children and less income-generating capacity, thus increasing the likelihood of the families' remaining in, or falling into, poverty and misery. The region needs to make an effort to identify alternatives for adolescents by increasing coverage, improving quality, and diversifying opportunities for receiving occupational training upon the conclusion of their basic education.

The potential of the education sector

The Secretariat called the countries' attention to the high profitability of a timely investment in education, the high rate of return on such an investment (estimated at 27 percent), and its effect not only on production but on the levels of health and quality of life of the population as well.

In addition, in view of the fact that educational institutions enjoy a "captive" population—which represents a unique opportunity to learn more about the living conditions, family and social environment, occupations, and health level of each student, as well as to form an integrated vision of that student—the Secretariat called for the sector to play a more active role in developing such knowledge, to strengthen its ties and coordination with other sectors, and to design mechanisms for counseling, orienting, screening, and referring students and their families in accordance with their particular conditions.

THE SANTIAGO AGREEMENT

In August 1996, the Government of Chile, with support provided by the Secretariat Pro Tempore and UNICEF, hosted the Third American Ministerial Meeting on Children and Social Policy. The objectives of that meeting were established jointly by the organizing entities and were defined as follows: (a) to analyze progress achieved at the midpoint of the decade toward compliance with the commitments made in the World Summit for Children and the Nariño Pact; (b) to reaffirm the commitment of the hemispheric countries to improving living conditions for children through a review of regional-level goals for the year 2000 and to subsequently establish new challenges and strategies; and (c) to promote the exchange of experiences among the countries of the region and international cooperation agencies and strengthen mechanisms of horizontal cooperation.

The status report compiled by the Secretariat served as the basis for discussions about achievements recorded, obstacles encountered, and challenges faced by the countries during the process. At the conclusion of the meeting, the countries signed the Santiago Agreement, in which they reaffirmed their commitment to:

- Develop, within the framework of the International Convention on the Rights of the Child, a systemic, integral, and progressive policy based on the rights and responsibilities of all actors. This policy calls for the development of social subjects (rather than objects of social program treatment), efficient participation in national policy making, and the production and distribution of social goods and services.
- Prevent and address emerging problems such as AIDS, substance abuse, early pregnancy, abandonment, sexual abuse, child labor, and violence.
- Ensure universal compliance with, and sustainability of, the goals of well-being for children and adolescents by the year 2000, through significant additional efforts aimed at accelerating, expanding, and increasing efficiency, effectiveness, and equity in social policy.
- Promote and support programs aimed at providing support to families.
- Provide a significant thrust to investment in human resource development with a view toward increasing and maintaining the quality of life of the people of the hemisphere.
- Strengthen the processes of democratization and the consolidation of a culture of cooperation in young generations and education in the habits and attitudes of solidarity, civic values, and human rights.
- Strengthen social information systems based on periodic mechanisms for gathering information, including principal social indicators, and encourage their use in social management by improving their availability, reliability, timeliness, comparability, and degree of disaggregation.

Specifically in the field of education, the Agreement establishes three primary goals deriving from the World Summit but adapted to the Region, together with fourteen auxiliary goals. The principal goals are as follows: to achieve universal access to primary education with equity in terms of gender, geography, ethnicity, socio-economic level,

and special needs groups; to increase to more than 80 and 70 percent respectively the percentages of children finishing fourth grade and primary school, and significantly increase access to secondary education; and to reduce by half the adult illiteracy rate, with special emphasis given to literacy programs targeting women.

The auxiliary goals identified in the Agreement are as follows:

- Establish systems for generating information on initial education and education for parents in order to ensure appropriate monitoring and evaluation of programs in these fields.
- Legislate, establish policy and regulations for improving initial/preschool education, and increase budgetary allotments to this sector.
- Promote the extension and improvement of family and community-based initial/preschool education programs.
- Promote the extension of parent education programs in the area of child development, with emphasis on the responsibilities of the parent in rearing the child.
- Reduce by half the rates of grade repetition in the initial grades of primary school.
- Identify strategies for decreasing the school drop-out rate.
- Increase levels of understanding in the areas of reading, writing, and arithmetic and increase the number of children completing fourth grade.
- Increase the equity and quality of basic education by introducing changes as follows:
  –on the management plane, by strengthening the administration, planning, implementation and supervision of the educational process and promoting increased participation by students, parents and community;
  –on the curricular plane, by adapting the curriculum to the needs of the individual, the community, and society and by promoting student-focused individual and group participative methodologies, expanding the role of the teacher as facilitator of the learning process, and making available appropriate materials to students; and
  –on the investment plane, by placing special emphasis on equity, through the allotment of sufficient infrastructure resources, professional training and development of teachers, texts, and length of the school day.
- Ensure the organization and use of systems for measuring the level and quality of learning.
- Develop flexible options for the education and technical-vocational training of young people, particularly business and computer courses.
- Include in the curriculum—as well as in the teaching materials and methods used in all preschool, primary and secondary education facilities—education on human rights, beginning with the rights of boys, girls, and women. This education is to be conducted in addition to education in skills for life, the environment, nutrition, and health, including reproductive health.
- Review and identify gender biases and any other type of discrimination in all preschool, primary, and secondary school teaching materials and methods; take corrective measures; and ensure the inclusion of notions of gender and social equity in new teaching materials and methodologies.

- Implement and expand programs and actions aimed at promoting the status of girls and adolescent women and mutual respect between boys and girls in preschool, primary, and secondary education facilities.
- Develop rehabilitation programs in support of children with disabilities and their families.

CONTRIBUTION OF THE SECRETARIAT PRO TEMPORE

There were a number of lessons learned during the first two years of operation of the Secretariat Pro Tempore. Most of the direct achievements of a network of exchange and monitoring such as that promoted by the Secretariat Pro Tempore are general, symbolic, and political in nature. The Secretariat succeeded in maintaining the interest and focus of social policies on children by contributing to the process by which national policies are made and increasing the awareness of governments vis-à-vis the need to evaluate their human development programs and improve their information systems. In addition, the Secretariat worked to develop consensus among governments, international agencies, and academic organizations with regard to the evaluation of common objectives.

National focus on children

The Secretariat has served as a mechanism that has enabled countries to maintain their focus on the commitments assumed by the governments of the hemisphere with regard to children, despite the changes in government administration occurring in each. During the period from 1994 to 1996, more than fifteen countries experienced changes in their heads of state, resulting in modifications to the national development plans of several countries, changes in the direction of social policy, and replacements in the management personnel of government institutions. The Secretariat monitors the agreements on children and acts as a showcase for the commitments, while working to ensure the continued focus of social policy on the guarantee, defense, and promotion of the rights of children.[7]

In this sense it may be said that the Secretariat functions as an adjunct to policy designers and implementers by working to ensure that their decisions and actions will contribute to making children's rights a reality and generate medium and long-term benefits for children.

Legitimatization of commitments and goals at the highest level

Countries are able to spearhead and promote ministerial meetings with support and advisory assistance from the Secretariat Pro Tempore and international agencies. These meetings constitute an opportunity to focus government attention on compliance with their commitments to children in their respective countries. At the same time, the signing of regional-level agreements makes it possible to review, define, and legitimize, at the highest policy level, the specified goals.

Adaptation of goals

The goals defined in the World Summit for Children are, as a result of their universal nature, indicative for the hemisphere. The meetings convened by the countries of the Americas, with support provided by the Secretariat and international agencies, have made it possible to clearly state the goals and even to define new goals for certain age subgroups not included in the world agreements.

In the case of education, it is possible to observe support for initial education and flexible and technical education for adolescents, as well as for the equity and quality of basic education beyond actual coverage.[8]

Improvement and use of information

The technical meeting, prompted by the Secretariat, pointed out the importance of monitoring goals while showing how the information systems developed by the countries did not respond precisely to the formulation of National Plans of Action for Children. In effect, possibilities for objectively evaluating social and child-related performance were limited, both at the level of each country as well as for the region as a whole.

A review of the social information systems related to such Plans' programs and activities revealed the multiple difficulties surrounding data production and analysis and in particular how that information was used. Also, it identified the widespread dispersion of sources and indicators as a significant constraint in consolidating the information required for monitoring the commitments.

Additionally, the countries reaffirmed the need to have access to information for designing the most appropriate policies for each country. With appropriate information, actions could be targeted with greater effectiveness, the intensity of specific interventions could be reinforced or decreased as needed, and the required financial resources could be procured and allotted. Information would also make it possible to identify progress made, achievements recorded, and obstacles encountered with regard to compliance with the proposed goals.

The decentralization process was identified as a significant opportunity to focus national efforts on improving records and involving the various social actors in the generation and use of the required information. In addition, it was shown that improving information mechanisms requires horizontal cooperation among countries in addition to the international technical support of agencies and other assistance providers.

Sensitization with regard to the problems of child development and protection

The International Convention on the Rights of the Child includes all the rights of children. The Plan of Action, however, focuses on problems of child survival. Although serious problems still remain with respect to child survival throughout the hemisphere,

there can be no doubt that other rights must also be promoted. The role played by the Secretariat and by international cooperation agencies has involved helping countries determine which of these rights should be prioritized, a process that has served to increasingly fine-tune the right to development and protection.

Analysis of child-related issues

The status report presented by the Secretariat during the Santiago Meeting is a clear demonstration of the potential of such a mechanism. The ability to access, on a timely basis, information available from governmental agencies in the various countries, international cooperation agencies, and other sources made it possible to organize a presentation and prepare a document that gives countries feedback about areas in which they have recorded the greatest achievements and encountered the greatest difficulties. In addition, the report was able to suggest specific areas toward which joint efforts should be directed.

The status report also identified inadequately explored issues, deviations, and emerging themes.[9] In this regard, it was recommended that the countries and agencies better focus their technical and financial efforts.

Exchange of successful experiences and methodologies

One of the activities carried out by the Secretariat, primarily in conjunction with its preparations for meetings, but also during the meetings themselves, involved identifying successful country experiences and methodologies in various areas related to children's issues. The exchange of such experiences and methodologies often constitutes a contribution to the policy design and program implementation processes, since an awareness of what others have done can aid in visualizing the potential of similar methodologies and experiences as they might apply to one's own country or region.

There are a number of advantages to hosting a Secretariat of this type. In Colombia, for example, such advantages included the strengthening of professional capabilities and the assurance of a continuation of the interest in children's issues at the highest levels of the central government, particularly among planners. Additionally, the ongoing close relationship with international cooperation agencies has served to strengthen the country's ties with such agencies. An additional benefit is the training received by those responsible for providing guidance and coordination to the Secretariat. A joint examination always serves to broaden horizons, provide instruction on issues not previously addressed, and ensure the application of current information both to the existing situation as well as to responses aimed at guaranteeing the rights of children.

The Secretariat and its relationship with the international community

In its role as an entity created by the governments of the countries of the Americas, the Secretariat performs an essential monitoring function. As a coordinating entity, it is responsible for enlisting the support of cooperative international organizations and

institutions with a view toward facilitating its operation. In addition, it is empowered to establish channels of communication and coordination with intraregional or international organizations or other entities working in the areas of children's issues and social development.

During its initial period of operation, the Secretariat established relations with international cooperation agencies through UNICEF, in deference to the role of the latter as interagency coordinator. At present this task is performed by the United Nations Fund for Population Activities (UNFPA).

The relationship with UNICEF was particularly fruitful, as that organization has in place a significant infrastructure at the regional and country levels, which it made available to the Secretariat. This led to much greater efficiency in the work performed and helped avoid duplication of efforts, which would have generated a high cost for any government.

Based on the Colombian experience, the governments decided assign the Secretariat the task of strengthening relations with the Interagency Coordinating Committee and its ties with UNICEF as an agency specializing in issues related to childhood and adolescence.

Constraints and obstacles to monitoring the agreements and goals

As previously mentioned, the primary obstacle to the monitoring of the agreements and goals in support of children is the lack of valid social information systems in most countries. There is considerable heterogeneousness in the definition of indicators, and information needed to document many of the goals is lacking.

Another type of obstacle results from the large number of changes occurring among technical and management personnel in the various countries. All too frequently it becomes necessary to repeat briefing information with regard to the status of international agreements and goals. Although the ministerial meetings have occurred at a high policy level, representatives as a rule are from the social areas of their respective countries (ministers of health, education, family, labor and human development), and compliance with the goals at the country level is not exclusively dependent on such individuals. In this regard, it would be desirable to focus efforts on bringing about the further involvement of the ministers responsible for the areas of planning and finance. One constraint involved with this mechanism is that it has not been able to incorporate key participants in civil society, among which boys and girls should play a leading role, into the monitoring process.

Key factors to bear in mind with regard to similar international mechanisms

The proper operation of a Secretariat Pro Tempore requires the countries' political backing. The task of monitoring agreements pertaining to children in the region that has been mandated to the current Secretariat grew out of, and is in turn strengthened by, the Nariño Pact and the Santiago Agreement. This is *a sine qua non* requirement

without which the Secretariat would surely be unable to operate. The country that assumes the role of host should have available qualified personnel to assume responsibility for ensuring a more dynamic Secretariat. In addition to technical capabilities, such individuals need to have extensive contacts with the academic community and with policy-making organizations. In addition to personnel, there is also a need for physical space and broad logistical support for purposes of communication and for international travel and dissemination activities.

The knowledge, available information, and infrastructure of an international cooperation agency, such as in the case of UNICEF with regard to children, are indispensable for ensuring the appropriate and efficient operation of the Secretariat. Without them, there would always be the risk that efforts might be duplicated or that activities might be pursued along parallel or misguided lines of action, in addition to which the possibility for increased outreach to countries and agencies would be lost.

Commitments and goals must be subjected to an intense process of debate prior to their eventual incorporation into ministerial agreements. Hence, there is the need to obtain the prior consensus of technicians from country governments and international agencies with regard to the viability of the goals. In addition, draft agreements need to be submitted for diplomatic review, or many countries may balk at signing them.

This process of ongoing discussion leads to social learning that can be observed, for example, in the increasingly precise fine-tuning of the goals, thus facilitating their monitoring and evaluation. As a case in point, the goals initially formulated were less precise than the ones eventually approved in Santiago. Nevertheless, compliance with the new auxiliary goals will not be easy. To achieve them, the countries will require support from a variety of sources as well as increased efficiency in their own educational systems. In this regard, it is recommended that organizations charged with international cooperation be fully aware of the goals and ascertain from the individual countries their current situation with regard to such goals, so that technical and financial assistance can be prioritized.

Additionally, specific meetings have been and could still be held to review child-related issues included in the agreements. A specific meeting comes to mind that promoted by the Secretariat with support from agencies specializing in education and children's issues and attended by the individuals responsible for policy making and implementation of educational programs. Such meetings make it possible to analyze achievements and constraints and identify alternatives for promoting attainment of the proposed goals.

In addition, efforts should be made to promote the participation of other players, such as school directors, teachers, and parents. It will not be possible to attain any of the goals if these players do not become involved on the participative and decision-making levels. National and international nongovernmental organizations can play a lead role in coordinating the members of civil society, not only for purposes of decision making, technical support, and the contribution of successful models, but also for the performance of tasks and the attainment of the stated goals.

The participation of children, both schooled and unschooled, is deserving of special attention. Children have a right to education and to participate in an authentic and nonsymbolic way in the decisions affecting them. For this reason, the principal recommendation with regard to the evaluation of the goals is to include the opinions of children in the evaluation process and, of course, in the formulation and implementation of policies and programs.

Future of the mechanism for monitoring the commitments to children in the Americas

The Secretariat Pro Tempore will continue to exist when the countries of the Americas so decide. In the opinion of the author, such a decision will be based on the perceived usefulness of the Secretariat to each government. Each country that hosts the Secretariat will impart to it a distinctive direction while conserving its fundamental mandate, which is to promote the exchange of successful experiences, define a regional agenda for horizontal cooperation, support and strengthen information systems, and periodically evaluate compliance with goals.

Probably one of the most effective ways to strengthen the Secretariat it to seek support through the establishment of partnerships and coordinating units with organizations and groups that share the philosophy of the International Convention on the Rights of the Child. Such organizations—with their considerable experience and authority in children's issues—can be fundamentally important for ensuring the participation of civil society, including children themselves, in the practice and monitoring of children's rights and of the goals defined by the governments to secure those rights.

In addition, close coordination with the Committee on the Rights of the Child and with the agencies of the United Nations and the Organization of American States in the form of an international consortium will assure proper monitoring of the Convention, which is the starting point for the agreements.

NOTES

[1] The nations that have yet to ratify the Convention are Somalia, the Cook Islands, and the United States of America.

[2] By late 1996, twenty-three of the thirty-four nations had already developed their respective NPAs, nine were in the process of preparing plans, and two had drafted working documents not yet accompanied by a second, policy-level document.

[3] The actions of the Secretariat Pro Tempore are analyzed only for the period from 1994 to 1996, since responsibility for the organization was handed over by the Government of Colombia to the Government of Chile in August 1996, and the author did not have available sufficient information on activities carried out subsequent to that date.

[4] The Committee is made up of the following agencies: the Pan American Health Organization (PAHO/WHO); the Food and Agricultural Organization of the United Nations (FAO); the United Nations Fund for Population Activities (UNFPA); the United Nations Children's Fund (UNICEF); the World Bank (IBRD); the Interamerican Development Bank (IDB); the International Labor Organization (ILO); the United Nations Educational, Scientific and Cultural Organization (UNESCO); the United States Agency for International Development (USAID); and the United Nations Development Programme (UNDP).

[5] The countries that reported having a validated scale were Chile and Colombia. Some countries (Peru and Costa Rica) reported that they were developing such a scale but that it was not yet in operation.

[6] According to information provided by UNESCO-OREAL, pre-primary school coverage for the population group between zero and five years of age reached a level of 14 percent in 1989.

[7] The Secretariat carried out this function primarily through informational communications, by sending packets containing the Agreements to the new governments, requesting information as to who would be in charge of monitoring the commitments, and soliciting data on progress achieved toward compliance with those commitments. The large number of countries responding to the Secretariat's requests for information, as well as their participation in meetings convened by the Secretariat, served to confirm this function.

[8] A review of the various agreements that have been signed makes it possible to infer this contribution by the Secretariat. It should be clarified, however, that the Secretariat did not perform this function independently, but rather jointly with the specialized agencies of the United Nations and the countries themselves. Although the adaptation of goals does not mean that the countries begin to comply with them immediately, it does mean that they begin to think about them and take them into consideration in designing national and local policies and programs. This statement can be substantiated through a review of the various National Development Plans prepared by the countries. One example is the inclusion in such plans of specific programs designed to increase the coverage of initial education, prevent child abuse, address disabilities, eliminate child labor, and modify legislation to make it consistent with the International Convention.

[9] The themes identified in the status report as emerging were HIV/AIDS, violence against children, child labor and exploitation, and reproductive health in adolescence.

Section II
LESSONS OF HISTORY

CHAPTER 5

SOCIAL IMPACT OF EDUCATIONAL PERFORMANCE EVALUATION SYSTEMS: THE CASE OF CHILE

Erika Himmel

The trajectory of systems for evaluating academic achievement in Chile offers valuable lessons for countries of the Americas implementing national testing programs, in at least two direct policy-related aspects: institutional responsibility and the impact of evaluation systems on society. The Ministry of Education has coordinated its efforts with those of universities and research centers to strengthen the country's institutional capacity and human resources for developing and administering a variety of evaluation systems. This chapter, in addition to describing the operation of such systems, provides a frame of reference for discussing how information generated by evaluation systems is used. The case of Chile is analyzed. Finally, the author identifies critical factors related to the use of academic achievement evaluation results.

INTRODUCTION

This chapter describes the operation of education performance evaluation systems, presents a frame of reference for discussing the issue of the use of information furnished by evaluation systems, and includes a case study of the Chilean experience. Lastly, the author looks at key factors in the use of educational performance evaluation data.

The first section of the paper presents a conceptual framework for examining national educational performance evaluation systems and the use of information produced by these systems and applies this frame of reference to a study of such evaluation systems in Chile. More specifically, it presents examples that clearly trace the use of information produced by these evaluation systems. Two of the systems in question assess achievement at the primary education level. The third assesses achievement upon the completion of secondary school. The paper also pinpoints a number of factors that have facilitated or hampered the use of evaluation data and puts forth certain proposals to help chart the course of national education policy.

NATIONAL EVALUATION SYSTEMS AND USE OF EVALUATION DATA

The past ten years have witnessed wide-ranging debates over the quality of education in Latin America. This is understandable considering how, in previous decades, area countries had been focusing all their efforts on improving the coverage of their education systems, thereby relegating the issue of quality to a lower plane. Countries that have gradually resolved enrollment problems have come to realize that the expansion of their education systems was achieved at the cost of compromising the quality of educational services, with investment requirements far exceeding their economic prospects. Hence the renewed interest at policy making levels in quality considerations, prompted by an acknowledgment of the importance of human capital for economic development. Added to these factors is the financial aid consistently channeled into the education sector by international agencies such as the Interamerican Development Bank, Organization of American States, and World Bank, all of which are interested in seeing the impact of this aid on improving local education systems embodied in tangible results (Lockhead, 1991).

Thus, the need to assess student achievement becomes apparent once an education system has succeeded in providing equal opportunity in terms of access to education and certain questions arise in regard to the quality of the educational services offered. It is within this context that national evaluation systems are set up to supply information on the achievement of educational objectives, pinpoint variables inherent in and outside the system explaining differences in performance, help make informed predictions on how the system will function in the future, and furnish indicators in regard to the system's most enduring features.

In general, evaluation systems consist of the administration of examinations or tests designed to measure the achievement of educational objectives in key curriculum areas. From time to time, these tests are supplemented by questionnaires on selected variables likely to explain differences in performance. Results of evaluations are generally made available both to direct stakeholders in the education process, such as teachers and school principals, and to indirect stakeholders such as the media.

The most frequently cited objectives of evaluation systems are as follows (Lockhead, 1996; Greany and Kellaghan, 1996):

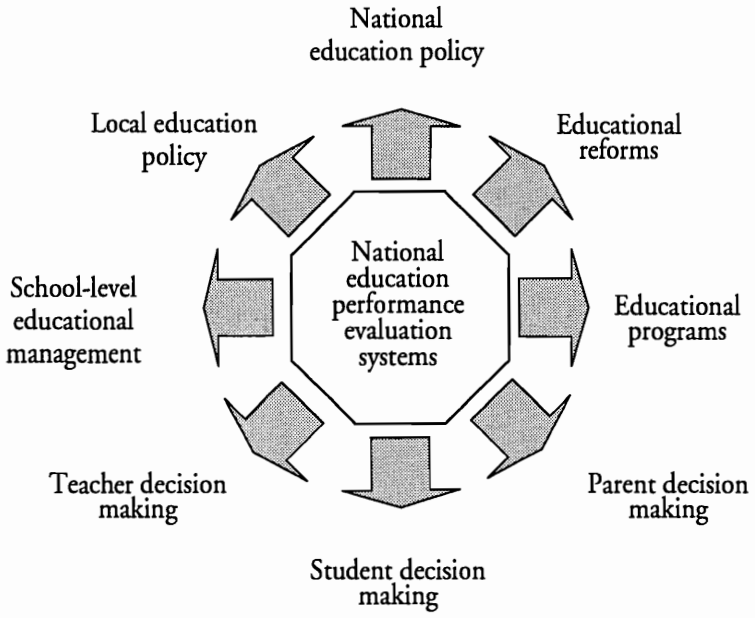
- to support and evaluate education policy;
- to evaluate specific educational programs;
- to monitor changes in educational achievement over time;
- to make educators accountable for student learning;
- to screen and place students moving onto higher levels of education;
- to corroborate student achievement;
- to furnish data to parents and their representatives on the quality of the education provided by the nation's schools; and
- to assess learning needs.

However, it is hard for a single evaluation system to achieve all these objectives, which explains why we often find two or more coexisting systems designed for different purposes.

In any event, an evaluation system with any of the aforesaid objectives should help provide a better insight into the workings of the education system, furnish necessary guidance for decision making by different stakeholders at different levels, and help improve the quality of educational services. In other words, the assumption is that, in addition to supplying information on the system, an evaluation of educational performance will actually affect the education system per se. The expectation is that this information will supply guidance for decision making, leading to different types of action whose effects can, in turn, be evaluated by these very same systems, provided they are still operative.

An educational performance evaluation system will affect different areas or groups, depending on its objectives, as illustrated in capsule form in Figure 1.

Figure 1. Selected Areas Impacted by National Educational Performance Evaluation Systems



Thus, if the objectives of a performance evaluation system are to produce information designed to back up and evaluate education policy and monitor corresponding educational achievements, its findings can have an impact on the propounding of new policies, including educational reforms. Moreover, they can also affect educational management at the individual school level and decisions taken by teachers with respect to the administration of the educational process.

However, these effects will materialize only to the extent that the evaluation data are actually used for decision making purposes, an assumption that is not always valid. According to a number of commentators (Alkin, 1979, 1985; Brown, Newman and Rivers, 1985), there are two approaches to conceptualizing the use of information, the "mainstream" and "alternative" approaches.

The mainstream approach conceives of the use of evaluation data in terms of their direct, immediate impact on the education system or program in question, or, in general, on the target of the evaluation. According to this concept, the use of data is regarded as an event, rather than as a process that can begin to take place as early as the planning phase of an evaluation system. This dichotomizes the dimensions of its use into two extremes, namely use versus nonuse. Adhesion to this mainstream approach means recognizing that an evaluation is being used only if and when it produces tangible effects such as the institution of short-term educational reforms, the replacement of one program by another, or major changes in educational strategies. These sorts of radical measures are not necessarily taken, since there is a series of factors that can influence the use of evaluation data which, in turn, are contingent upon the elements of the evaluation process and the findings produced by this process. Moreover, there are other factors taken into consideration in decision making processes, in addition to evaluation data.

In an attempt to further clarify this concept of use, King and Peachman (1984) maintain that it is based on certain questionable assumptions, as described below:

- Decisions can be made in a classically rational manner, without taking into account political, social, and organizational variables affecting decision making processes;
- Evaluation data are the only factor triggering immediate, observable effects (the myth of the "big bang" theory);
- The quality of evaluation reports alone suffices to guarantee their full and complete use; and
- Active cooperation between evaluation personnel and decision makers will automatically step up the use of corresponding data.

As a matter of fact, the administrators of the Performance Evaluation Program or PER conducted in Chile over the period between 1982 and 1984 began operating the program based on many of these assumptions, mistakenly presuming that resulting data would be used by teachers, school principals, and education officials to formulate proposals for the mounting of projects and programs at all levels via a self-management and self-monitoring process. However, they soon realized the weaknesses of these assumptions and embraced the alternative approach.

Another example of this concept of use is found in Schiefelbein (1992: 264), who judges the impact of Chile's PER on the sole basis of the finding that there were no significant changes in student performance on corresponding evaluation instruments,

commenting that "this shows that expectations in terms of improving academic achievement are too high."

On the other hand, in the alternative approach, the use of evaluation data is conceived of as a gradual process in which such data, along with other information, may eventually lead to the mounting of small-scale projects and programs that, little by little, change the baseline situation. Under this concept of use, the impact of the evaluation is far from immediate. It can take years before the impact materializes which, combined with other contextual information or under a different set of circumstances, means that it can take on different meanings at different points in time (Braskamp, 1982).

King and Peachman (1984) believe that there are at least three levels of use of evaluation data, namely: symbolic or suasive use; conceptual use; and instrumental use.

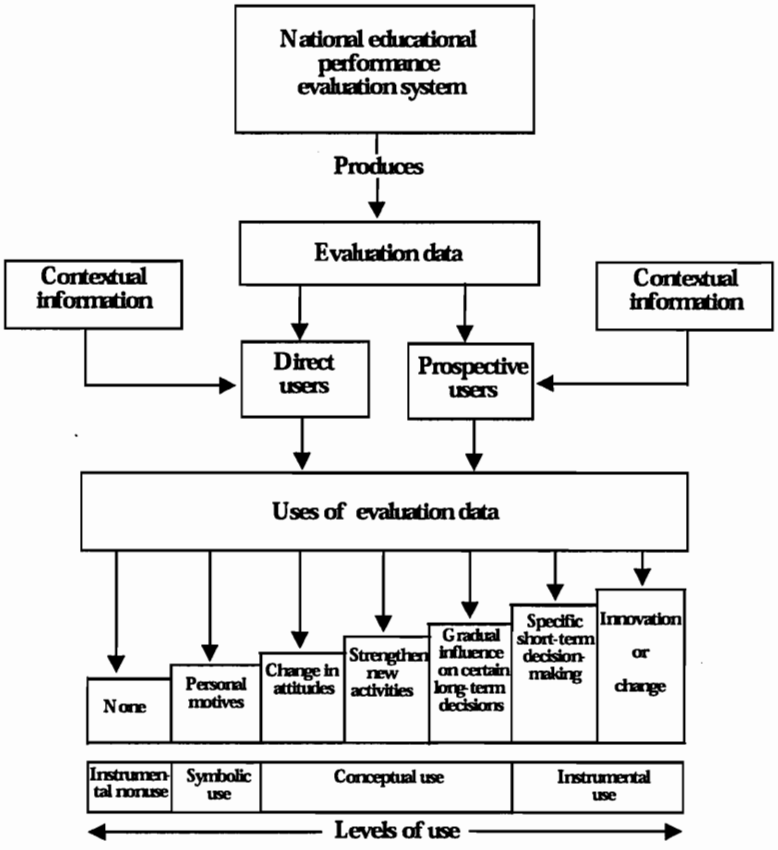
There is also a fourth level, that of instrumental non-use. The relationship between the production of evaluation data, contextual information, and the use of data as an input in decision making processes is outlined in Figure 2.

The symbolic or suasive use of evaluation data refers to its use for personal purposes. For example, the findings from an evaluation may be used by a school principal as evidence of good performance, providing facts and figures with which to secure his position. In such cases, a study of the specific context is needed to accurately ascertain the underlying intentions of the party. In this example, the evaluation is being used to justify certain decisions rather than as input in such decisions, and is often used in this sense for political or administrative purposes.

This first level of use includes a specific effect of performance evaluation systems that is extensively discussed throughout the literature (Greany and Kellaghan, 1996), namely their influence on teaching. In fact, no outside testing program is neutral, and teachers, somehow perceiving the importance of corresponding test results, will begin focusing the teaching process primarily on the content areas and objectives targeted by the testing instruments, and base their approach to the school curriculum on the coverage of these tests.

The second level of use, conceptual use, refers to the evaluation causing users to reflect upon the aim or purpose of the evaluation, prompting them to acknowledge the existence of certain developmental skills or problems, thereby triggering a change in attitude. Over the long run, this change in attitude may cause them to make certain highly specific decisions. For example, a school principal may attribute low student achievement on a performance evaluation to a lack of identification with the school. This conclusion may prompt him or her to organize a discussion session with the teaching staff to pinpoint the causes of this problem. In this example, while no specific action was taken, the evaluation data nevertheless encouraged the user to reflect on the problem, and eventually resulted in the mounting of small-scale efforts such as informal meetings of students and teachers.

Figure 2. National Educational Performance Evaluation System



Instrumental use, the third level, is characterized by clear linkages between the data produced by an evaluation system and certain decisions taken by a direct user of such data. Using the illustration presented in the previous paragraph, should the principal decide to systematically poll the student body as to the positive and negative aspects of the school for purposes of mounting a program designed to improve student identification with the teaching staff, this would be an example of the instrumental use of the information furnished by the evaluation system.

Lastly, instrumental non-use is where information is intentionally discarded by corresponding users. If, for example, the findings from a performance evaluation reveal that fourth-grade pupils in a particular municipality have performed poorly in spelling, and the education department of the city government in question feels that exact spelling is not a basic educational objective at that grade level, then it is likely that no effort will be made to improve student achievement with respect to this educational objective.

In analyzing the use of evaluation data according to this model, the effects of a national educational performance evaluation system may be reflected in many activities that do not necessarily produce short-term improvements in performance. However, over a medium or long-term time frame, such efforts may improve the quality of student learning.

For the purposes of this study, social impact refers to types of uses of evaluation data affecting action-oriented decision making with effects on or implications for social groups within or outside the education system. Moreover, the effect of the evaluation on the decision making process per se is barely perceivable, and at best is inferred from certain explicit signals in the ensuing form of action.

Educational performance evaluation systems in Chile

This section of the paper describes the country's three longest-running evaluation systems. The first, the Performance Evaluation Program, or PER, targets basic or primary education. It is included in this discussion because it served as the basis for the second such system, the National Educational Quality Assessment System, or SIMCE, which has been in use since 1988. The SIMCE extended the evaluation process to the second year of secondary education. The third system, which has been operative since 1967 and is administered to students completing their secondary education, is the National Admissions System for Higher Education (*Sistema Nacional de Ingreso a la Educación Universitaria*).

The Performance Evaluation Program (PER)

The issue of assessing educational quality was first broached by Chile's Ministry of Education in 1978, when it approached the Pontificia Universidad Católica to design and set up an educational information system. As part of the government decentralization process underway at the time, this program helped collect data on educational quality and disseminate it to different stakeholders in the education process for the mounting of efforts designed to raise the quality of education.

This task was entrusted to a multidisciplinary team of experts which, under the terms of a specific agreement, was charged both with conducting the corresponding feasibility study and with starting up the program. The team consisted of a core group of professionals on the University faculty (most of whom had done their graduate work in the United States), including four engineers, three psychologists, three educators, and one sociologist. The psychologists and educators were in charge of designing the various educational assessment instruments in cooperation with national education officials and education/administration system officials at the regional level. The engineers were responsible for program administration, logistics, and computations (Himmel, 1996).

The overall objective of this program was to help improve the quality of education through the decentralization of educational authority. The expectation of the team members was that all stakeholders would take a more active role in this process as a result of the performance evaluation, with the underlying assumption being that the mere fact of furnishing information on student achievement of basic educational objectives would prompt teachers and school principals to make different types of innovations designed to improve the quality of education. Moreover, the availability of objective, reliable, well-grounded data for national education officials was expected to produce a more realistic approach to education policy.

The specific objectives of the PER were the following:

- explicitly establish educational objectives at the primary school level whose achievement was considered essential by the Ministry of Education;
- furnish information to parents; teachers; school principals; and local, regional, and central government officials on the achievement of these educational objectives with appropriate levels of specificity and aggregation to meet the needs of each target audience;
- provide necessary information to the Ministry of Education for the fulfillment of its new policy making and supervisory functions, with such information designed to pinpoint low-scoring schools requiring greater technical/educational and financial assistance and to monitor the performance of individual schools and municipalities, and realign central planning with respect to curriculum development, in-service teacher training, and the design of textbooks and instructional materials.

The main elements of the program design were as follows (Himmel, 1996):

- The program measured educational achievement at two grade levels: the fourth grade, because it marked the end of the first stage of general basic education and the last year in which all subjects are taught by the same teacher, and the eighth grade, which was the last year of compulsory education in Chile, with subjects at this grade level taught by different teachers.
- With student skills and proficiency in language (reading comprehension and written expression) and mathematics (computation and problem-solving) regarded as the backbone of the primary education process, tests in each of these areas were administered to each student in full. The other assessment instruments, in the natural and social science areas and the areas of personal and social development, were administered by means of matrix sampling with each student responding to a sample of test questions, taking the class average to denote achievement in that particular area.
- With the exception of assessments of written expression, all tests were incremental, in the form of multiple-choice questions, covering all educational objectives at the level of education in question.
- Tests were constructed in accordance with stringent international standards by teachers, as well as by experts in curriculum development and evaluation, and focused on an array of basic educational objectives and minimum content areas. The final versions of all multiple-choice tests had a confidence level of over 0.90 for assessments in cognitive areas, and of over 0.80 for assessments of personal and social development. The writing test had an intercorrecting confidence level of over 0.80 with the analytical method used for purposes of this testing instrument.
- It was decided to administer the tests to the entire student population, on the assumption that stakeholders would be moved to action only by findings relating to their particular domain. Thus, only extremely small schools (schools with fewer than five students at each corresponding grade level) and schools located in remote areas (which were inaccessible during a large part of the year) were excluded from the testing process. The tests were administered to approximately 400,000 students per year, representing a student coverage rate of 90 percent.

- Schools were divided into homogeneous groups according to the socioeconomic status of their corresponding student body, based on criteria used in previous studies (Himmel et al., 1980; Himmel et al., 1982) to distinguish the effect of efforts undertaken at the school level from the influence of structural variables. These groups were referred to as "socioeconomic frameworks." An exhaustive study was made of the possible advantages of furnishing data on average student socioeconomic indicators at the school level along with corresponding test results, which was the procedure used in the United States (Alkin, 1981). However, it enabled a frame of reference facilitating more personal interpretations and analyses by target audiences in general, and by teachers and school principals in particular.
- Information activities undertaken in preparation for the administration of corresponding test instruments consisted of the distribution of leaflets to school principals, teachers, and parents explaining the program objectives. In addition, a technical pamphlet was distributed each year describing examples of the assessment instruments to be administered that year. Finally, administration system personnel distributed a series of audiovisual materials to local officials, school principals, and teachers.
- Tests were administered at the end of the school year under controlled conditions to ensure the homogeneity of corresponding testing procedures. To accomplish this, a country-wide administration system was set up, consisting of a network of 640 supervisors. A total of 12,000 examiners were given training, and extremely strict security measures were taken in regard to the tests.
- The test results were announced the following school year during the first month of classes, presented in the form of data on the average percentage of correct answers on each test and for each set of objectives. Reports were drawn up at different levels of aggregation for different target audiences. Reports for teachers, for example, furnished data by objectives and class totals. Reports for school principals presented corresponding data by school and grade level, while reports for other officials furnished data at the local, regional, or nationwide level, according to their respective sphere of authority. In addition to data presented at the aforesaid levels of aggregation, the reports also included two indicators for purposes of a comparative interpretation. The first consisted of the twenty-fifth and seventy-fifth percentiles of student test scores for schools within the same socioeconomic framework. The second consisted of the fifth and ninety-fifth percentiles of student test scores at the nationwide level. Thus, subsequent evaluations furnished comparisons with previous years.
- Manuals for the interpretation of the test results and audiovisual aids were prepared and distributed to national and local officials, school principals, and teachers.
- Specific reports were made to the Ministry of Education on schools urgently requiring technical and economic assistance.
- This information was rounded out by the preparation of teaching guides for those educational objectives with respect to which student achievement was poorest at the nationwide level. These were distributed to all teachers at the targeted grade levels.
- The ministry was presented with a series of recommendations in line with the program design for the channeling of technical/educational assistance to schools with the poorest performance ratings through school supervision.
- The program team introduced a series of changes in the educational evaluation system, in all cases verified by empirical data, particularly where the proposed

changes went against certain deeply-entrenched beliefs. For example, in a survey of teaching personnel, teachers maintained that fourth-grade pupils were incapable of answering test questions on separate answer sheets. The answer sheets were subsequently tested on a sample of pupils with an extremely low socioeconomic status. The results showed that there was no significant difference between the averages for pupils using an answer sheet and those whose answers were written directly in the test booklet.

The initial agreement entered into between the Ministry of Education and the Pontificia Universidad Católica de Chile for the performance of the feasibility study and start-up of the PER was for one year. It was renewed for another three years for program implementation purposes. Upon its expiration, the agreement was not renewed by the Ministry of Education, thereby terminating the PER.

Although it was impossible to detect a significant improvement in the achievement of educational objectives within this brief period, there were, nevertheless, substantial gains in the achievement of certain objectives in which student performance was extremely poor on the initial assessment as a result of remedial measures taken by teachers and schools. For example, there was an improvement of nearly 10 percent in verbal problem-solving in the mathematics area, at both grade levels targeted by the evaluation.

The country had a string of six different education ministers, with resulting replacements of top-ranking ministry officials, throughout the course of the program implementation period, the latest of whom announced his intention of terminating the PER in early 1984, the final year of the agreement with the Pontificia Universidad Católica. This decision was never discussed with the project team, and the reasoning behind it is unclear; all attempts at negotiation on the part of Ministry officials supporting the program ended in failure.

It is safe to say that one of the determining factors in the decision to terminate the PER was its cost of US$5.00 per student, or roughly two million dollars which, with the country in the midst of a recession, may have overwhelmed government officials. A follow-up study on the PER (Himmel et al., 1988) revealed that, at the time, top-level government officials were not committed to the program and felt that its cost was too high.

Another guess is that there were opposing factions within the Ministry of Education bureaucracy, one in favor of and one against the program being run by the Pontificia Universidad Católica. Those against could have been members of groups within the Ministry that considered themselves capable of implementing the project.

Another plausible explanation lies in the general political climate, which at the time was dominated by a group of economists advocating neoliberal economics, while a large percentage of the teaching profession clung to the notion of "state" education, or the vesting of all educational authority in the national government. In addition, government decentralization efforts were virtually at a standstill at that time.

A follow-up study of programs and projects mounted during the course of the operation of the PER (Himmel et al., 1988) (conducted upon the expiration of the agreement), found that certain schools had taken specific measures designed to improve student learning, such as exhaustive studies of syllabi to establish educational priorities and reviews of evaluation forms used by teachers. These efforts produced a change in the "evaluation culture" and prompted the implementation of a series of administrative measures such as an increase in the number of instructional hours devoted to language and mathematics (Contreras, 1988).

The National Educational Quality Assessment System (SIMCE)

By 1988, Chile's education system had been completely decentralized. However, there was still a large faction within the Ministry of Education that felt it had the technical expertise to develop a student performance evaluation system if vested with the necessary economic resources. The authorities felt that the reinstatement of an evaluation system not only required additional funding, but that it basically required a technical expertise that the Ministry did not possess.

High-ranking Ministry officials recommended the following measures for reinstating the evaluation system, ensuring its political acceptability, and reducing its cost (Himmel, 1996):

- Change the name of the evaluation system to disassociate it from the PER: hence, the National Educational Quality Assessment System, or SIMCE.
- Enter into new agreement with the Pontificia Universidad Católica, entrusting it with most aspects of project implementation for three years, on condition that it agree to train a team of Ministry of Education officials to take over the SIMCE upon the expiration of the agreement. The Ministry-based team to assume full responsibility for project implementation during the fourth year of the implementation period, with advisory assistance from the university-based team. When the agreement expires, institutionalize SIMCE within the framework of the Ministry of Education.
- Administer assessment instruments to the total student population, focusing on language and mathematics. Put the university in charge of the technical aspects of test administration based on a model similar to that of the PER.
- Administer science, history, and geography tests to a sample of 10 percent of the student population, with corresponding test instruments to be designed by the university-based team.
- Make an independent team within the Ministry of Education responsible for assessing personal development, as well as for including assessments of student, parent, and teacher satisfaction with educational services and of selected indicators of educational efficiency, such as repeater, promotion, and dropout rates.
- Administer assessment instruments at the fourth and eighth-grade levels, in alternate years.
- Put the ministry in charge of the process of administering the testing instruments or, in other words, of setting up the administration system and hiring examination personnel, while making the university responsible for system logistics and computations.

- Use the university-based team to undertake information activities designed to publicize the system, heighten public receptiveness, and disseminate corresponding findings.

The SIMCE was to have the following objectives:

- Assist the Ministry of Education in its policy making role and in over-seeing the education system;
- Bolster supervisory efforts by regional and local officials and provide them with technical support and assistance;
- Assess the quality of each educational institution, make comparisons, pinpoint explanatory factors, and evaluate educational program performance;
- Provide guidance for in-service teacher training activities, oversight efforts, and resource allocation.

Note the similarities between the objectives of the PER and those of the SIMCE.

Since its inception, the system has been running smoothly, subject to certain changes introduced in nonacademic areas such as assessments of creativity and personal development (Prado, 1995).

One of the current and possibly unique features of this system is its publication of test results at the country-wide, regional, local, and individual school levels in the nation's press since 1995, which has helped heighten public awareness of the SIMCE.

Moreover, the SIMCE is already regarded as an institution, forming an integral part of the regular activities conducted by the nation's schools. School principals and teachers alike recognize the system's sound technical features and regard it as a valuable educational management tool, although they maintain that it does not provide a complete picture of educational quality (Zabalza et al., 1994).

The National Admissions System for Higher Education

Until 1966, a "bachillerato"—the Chilean version of the French baccalaureate extended to students passing an examination administered by the University of Chile since 1927—was required by law for admission to the country's eight universities. In 1966, the Chilean Congress abolished this requirement, forcing the country's universities to design a placement system.

The University of Chile had developed and tested a preliminary version of a placement system as far back as 1962, which consisted of a series of objective, multiple-choice tests combined with consideration of the applicant's secondary school grades (Lara, 1985).

The abolishment of the "bachillerato" and the need to screen applicants for admission to the nation's universities within a very short time frame prompted the formation of an intercollegiate ad hoc committee vested with broad-based powers to tackle this issue.

The committee charged the University of Chile with the task of administering the placement system, given its technical experience in this area. This ad hoc committee was later reorganized into what is currently known as the Chilean Board of University Chancellors Admissions System Coordination Committee, whose main function is to assure the due and proper operation of the placement process.

The main features of this admissions system, which has changed very little over the thirty years it has been in existence, are as follows:

- Until 1982, placement tests were required for admission to all eight universities existing at that time. These tests are presently required only for admission to government-funded universities, that is to say, to the country's eight original universities and their offshoots. Many private universities receiving no government funding do not require these examinations.
- The placement system consists of a series of mandatory multiple-choice tests and a second series of elective tests in the same multiple-choice format, from which applicants can either take whatever tests are required for their intended field of study or take all the tests. The tests are scheduled to avoid any overlapping. One of the mandatory tests is the Academic Aptitude Test, or PAA, which is similar to the SATs administered in the United States, consisting of a verbal and a mathematics portion.
- The other test that is currently mandatory for all applicants is a test on Chilean history, which was added in 1984 to promote the teaching of Chilean history at the secondary school level.
- The series of elective tests, known as specific achievement tests, meet the needs of specific university programs. These tests cover five secondary school subjects: mathematics, biology, physics, chemistry, and the social sciences. For a time, there was a single science test for biology, physics, and chemistry.
- All tests are scored on a standard scale. The average test score is 500, with a standard deviation of 100.
- The other type of data used as a screening criterion are the applicants' average secondary school grades, expressed in terms of the same scale used to score the placement tests.
- The tests are administered at the end of each school year. Students register to take the tests approximately three months in advance. Six weeks after the administration of the test, the results are published in a newspaper with a nationwide circulation using each applicant's national identification card number.
- Once the test results are published, students can apply to any of the country's traditional universities or their offshoots in a single application process and may apply for admission to several programs at different universities.
- The placement process per se is conducted by the University of Chile using the criteria and weights established by each university. Most universities set criteria and assign weights based on independent studies of predictive value. The only stipulation is that they assign a weight of 10 percent to the Chilean history test and a weight of 20 percent to applicants' secondary school grades. The results of this placement process are, in turn, published in a newspaper with a nationwide circulation by university, broken down by study program.

- The entire process is coordinated by the aforementioned coordination committee.

This system has been used for the past thirty years and, despite periodic demands for radical changes in the system, no alternatives have been put forward to date that can possibly hope to compete with the current system in terms of efficiency or cost.

Social impact of educational performance evaluation systems in Chile

Social impact of the SIMCE on education policy

The following section of the paper discusses three initiatives mounted in accordance with Chilean education policy guidelines, grounded in findings produced by the SIMCE, namely the "900 Schools Program" (P-900), the Program for the Improvement of Rural Schools (MECE-RURAL), and the Educational Advancement Projects (PME) initiative. These initiatives are all examples of the instrumental use of data supplied by the SIMCE.

The 900 Schools Program

This program, targeted at schools in impoverished areas, was the first implemented by the country's new democratic government in 1990. It was directed at the 10 percent of the nation's schools showing the poorest performance under the SIMCE, all of which serve a student population living in abject poverty. This assessment was made based on SIMCE data for 1988 and was said to be a form of positive discrimination, emphasizing student achievement at the first and fourth grade levels in the areas of language and mathematics. The program was also designed to promote student creativity, as well as personal and social development skills. It also trains teachers in methods of making education more relevant to their students' cultural environment and of promoting cooperation between the school and the community.

The program is divided into the following components:

- Participatory, inservice teacher training workshops designed to update academic training, and inspire thought and reflection on teaching practices with a view to promoting innovations in classroom techniques and establishing professional standards for addressing the issue of student achievement.
- Student learning workshops, known as TAPs, giving special attention to third and fourth-grade pupils requiring remedial instruction, run by young extension workers between 18 and 25 years of age from the surrounding community. They are given regular training in all areas of their work as well as with the children with whom they are experiencing any problems (Cabezón, Condemarín, and Vaccaro, 1996). These workshops are conducted outside regular school hours.
- Supplies of educational materials, including classroom libraries and teaching materials, with an emphasis on games, to bolster teaching efforts in the areas of language and mathematics at schools assisted under this program. Among the teaching materials distributed through this program were large numbers of pocket calculators.

- Infrastructure improvements, consisting of the repair and upgrading of those elements of the school facility regarded by teachers as most capable of bolstering student learning.
- Technical/educational supervision, administered by supervisors attached to the Ministry of Education, who are, in turn, given job training under the P-900 Program.

The program is open to schools with SIMCE test results well below the region-wide average. The schools graduate from the program as soon as corresponding evaluation data falls in line with the regional average or when improvements in their performance outstrip those of other public schools. This means that SIMCE data are not only used to determine the eligibility of schools for acceptance into the P-900 program, but also serve as a yardstick for assessing the program's impact. The number of schools enrolled in the program jumped from 900 in 1990 to roughly 1,300 by 1991, subsequently falling back down to approximately 1,000 by 1996. Fluctuations in the number of schools enrolled in the program are a raw indicator of the improvement shown by a selected group of schools in that it implies the attainment of better SIMCE test results. Close to 14,000 youths have been trained as community extension workers under the program. Through this training, these youths—who come from the same poor backgrounds as the students served by the schools in question—have improved their basic education and social skills.

Program for the Improvement of Rural Schools (MECE-RURAL)

This program is an integral part of an initiative mounted under the Program for the Improvement of Quality and Equity in Education (MECE). It is designed to improve the quality of educational services offered in schools with high educational risks located in rural areas, which house 20 percent of the nation's population and whose diversity stems from the country's geographic features. A large share of the rural population is concentrated in two areas of southern Chile that have performed extremely poorly in nationwide evaluations. A pilot project conducted in 1992 as part of the P-900 initiative led to the implementation of the MECE-RURAL program in 1993 targeting one, two, and three-room schoolhouses (Parraguez, 1993).

The main components of this program are as follows:

- the proposal of new educational methods in the form of a curriculum development manual. The manual related general skills and knowledge to the local cultural environment in an effort to make the school curriculum more relevant in each area of the country;
- teacher training for the implementation of educational reforms. The training would be designed to improve student learning in rural areas and make use of educational methods outlined above;
- development and supply of first and sixth-grade textbooks designed for rural settings, in line with the proposed new educational methods;
- development and supply of teaching materials for students and teachers to help bolster learning in the areas of language, mathematics, and science;

- classroom construction and hiring of teachers to expand schools that do not presently go as far as the eighth grade, which is the last year of compulsory education in Chile;
- establishment of "rural education coordination microcenters" for the organization of all teachers in schools within a single geographic area, providing these teachers with a regular forum at which to discuss their teaching practices, share previous and create new experiences, evaluate their progress, and obtain technical/educational assistance from supervisory personnel; and
- as in the case of the P-900 program, a strong technical/educational supervision component providing corresponding supervisory training,

The percentage of one, two, and three-room schoolhouses covered by the SIMCE was comparatively low until 1996, when the number jumped abruptly. A report published by the Ministry of Education claims that fourth-grade test scores in schools participating in the MECE-RURAL program rose six points in language and eight points in mathematics over the period between 1994 and 1996 (Ministry of Education, 1997), explaining that, strictly speaking, this claim was not grounded in an empirical assessment, complete with control groups. However, the tests used in the MECE-RURAL program are the same as those administered at the country-wide level. SIMCE data is being used to pinpoint problems and will eventually be used to monitor programs designed to overcome these problems.

The Educational Advancement Projects (PME) Initiative

This initiative was designed under the MECE program as a general strategy for the decentralization of education in Chile, a country whose education system has traditionally been highly centralized. In its early stages, the initiative focused exclusively on decentralizing economic school management, ignoring the issue of the decentralization of educational management.

The objectives of this strategy are to improve learning among primary school students from poor backgrounds with a view to reducing repeater and drop-out rates. It calls for each school at the primary or basic general education level to design its own educational reform projects, preferably in language, mathematics, natural science, and social science, in an independent, participatory effort by the school's teaching staff.

Projects are awarded under a competitive bidding process open to virtually all basic education establishments operated by municipal governments and government-funded free private schools.[1] Only schools previously conducting educational advancement projects or having recently enrolled in the P-900 program or the MECE-RURAL program are barred from competing. One, two, and three-room rural schoolhouses may compete only through their respective "rural microcenters."

The PME initiative includes teacher training for the designing of educational advancement projects, project preparation and implementation, the ongoing improvement of instructional methods and content, the use of new teaching resources, and project evaluation.

Efforts by participating schools to prepare educational advancement projects are bolstered by advisory assistance furnished by technical/educational supervisors previously trained under the PME program. These advisory services consist of training conducted within the framework of "communal workshops" for schools within the same geographic area. The emphasis is on project preparation procedures, and leaves the issue of content aside to encourage each school to address its own problems.

Projects submitted for approval by participating schools are subject to a two-stage appraisal process based on clearly defined project appraisal criteria. In the first stage, conducted at the local level, the projects are appraised by a committee of supervisors trained by the MECE management team. In stage two, they are approved and awarded to the schools at the central level.

The projects are appraised from the standpoint of their quality, equity, and relevance. The quality of a prospective educational advancement project is judged by its potential to produce an improvement in the teaching and learning process and thus better student achievement. The equity factor favors schools with larger learning deficits as measured by average scores obtained by students on SIMCE language and mathematics tests. Relevance refers to how well the proposed educational advancement projects are attuned to the surrounding socioeconomic and cultural environment.

A part of the project design process at the individual school level, one of the main criteria for selecting the problem to be addressed by the project is an analysis of how the SIMCE data measures student achievement with respect to educational objectives. Another criterion used to evaluate the efficiency of this strategy is the extent of changes in SIMCE evaluation data that occur after the project is implemented.

Thus, this is a "bottom-up" strategy for improving the education system. Moreover, SIMCE data is used as a criterion both for awarding projects and for evaluating their success and, by teachers, as an input in targeting areas in which student achievement is weakest for the introduction of reforms. This use of SIMCE data is an example of how the impact of evaluation systems can be heightened at those levels of the education system directly involved in the delivery of educational services.

Social impact of the SIMCE on teachers and schools

Although there is relatively little research data on the social impact of the SIMCE on teachers and schools, two studies do address this matter (Zabalza et al., 1994; Cerda et al., 1996), as discussed below.

To begin with, there are discrepancies between rural and urban schools with respect to the administrative management of programs and projects grounded in SIMCE data, which can be explained by differences in school size and organizational structures. Thus, reform projects in rural schools are headed by the school principal, working with the teaching staff. On the other hand, projects in urban schools are managed by an intermediate-level unit within the school's organizational structure known as a techni-

cal/education unit, with very little or no input from the school principal. Managers of 29 percent of all urban schools admit taking no measures in response to SIMCE findings because this is a matter for the technical/education unit (Zabalza et al., 1996). This fact can limit the impact of new initiatives, whose success has been found to rely on effective leadership by the school principal (Rutherford, 1985; Davis and Thomas, 1989; Enrione, 1991).

Educational advancement efforts undertaken by schools in both rural and urban areas include the regular administration of tests similar to those administered under the SIMCE to assess student achievement. Other such efforts include the framing of remedial strategies and the monitoring of these initiatives. Other innovations include the institutionalization of the concept of devoting more instruction time to language and mathematics and the promotion of new teaching methods (Zabalza et al., 1994).

Rural schools are also putting more emphasis on the personal development of their students and, more specifically, on the development of traits such as self-reliance, a sense of responsibility, and self-esteem. School principals report students in rural schools to be weak in these areas, which becomes apparent when these students with their urban counterparts for admission to secondary schools (Zabalza et al., 1994).

There is a certain group of schools that study SIMCE data but do not use this data to make innovations, in most cases claiming a lack of funding. Nor do these schools show appreciable changes in their performance over time (Zabalza et al., 1994).

Efforts such as the implementation of educational advancement projects have helped create fora enabling teachers to discuss teaching methods and share experiences with their peers. These activities are highly regarded by participating teachers, who, nevertheless, report scheduling problems associated with their school work schedules, making it difficult to juggle the timing of these meetings (Cerda et al., 1996).

Teachers attribute improvements in performance under the SIMCE to factors such as innovative strategies, new teaching materials used as part of the teaching/learning process, and regular classroom evaluations. Likewise, poor performance is attributed to students (learning problems, poor study habits, weak cognitive skills), families (the provision of inadequate cultural models), and school management.

Social impact of the SIMCE on households

The impact of the SIMCE on individual households can be examined from two perspectives, that of the degree to which SIMCE data affect the choice of a school for the children of the household and the extent to which the data heighten parental influence and participation in schools.

We found no empirical data on the former factor. However, informal observations show that some parents have gradually come to consult SIMCE test results in making deci-

sions as to which school to send their children, particularly since this information has been given extensive coverage by the media. Obviously, only parents in the middle and higher-income brackets, who have the option of choosing between public and private free and pay schools, are using SIMCE data for this purpose. Parents in low-income households do not have these options; their only alternative is to send their children to nearby free schools which, in impoverished areas, are in extremely short supply.

In examining the latter factor, again, we found no evidence that a knowledge of SIMCE test results was directly responsible for heightening parent participation in and influence on the educational process. However, an evaluation of PMEs (Cerda et al., 1996) found that in schools involving parents in specific educational activities, parents who came to realize that their children were learning new skills as a result of the educational interaction between parents and their children began to take more interest in the school.

These are examples of the conceptual (data analysis with no further implications) and instrumental (data analysis followed up by specific measures) uses of SIMCE data. In sum, it appears that policies grounded in assessment systems implemented directly by the Ministry of Education still have a much greater impact than independent initiatives undertaken at the individual school level, which continue to be rather limited in scope. This could be a negative effect of the highly centralized nature of the country's school system, with individual schools still waiting for instructions from national headquarters before mounting any type of effort.

Effect of the National Admissions System for Higher Education on student decision making

The impact of the National Admissions System for Higher Education on decisions taken by students has changed over the years, largely due to structural changes within this education subsystem. In fact, up to 1981, the system of higher education afforded secondary school graduates the options of being admitted to a university, pursuing less reputable nonuniversity-based studies, or entering the job market. In practice, most graduates who were not immediately accepted into a university chose to retake the entrance examinations several times. After several tries,[2] these students were often admitted to a university, though not necessarily into the program of their choice, generally in a field of study with a lower social stature (Himmel and Maltes, 1980; Himmel and Maltes, 1987). This process had an impact both on students and their families in terms of time, economic resources, and frustration.

With the restructuring of the country's higher education system, large numbers of middle and high-income students have found a solution in private universities, which generally offer programs of study with high social stature and good job prospects (Muga and Rojas, 1993).

On the other hand, these private universities are unaffordable for secondary school graduates from poor backgrounds, to whom institutions of higher learning other than universities offer only piecemeal solutions.

Unforeseen effects of educational performance evaluation systems

The following is a discussion of selected factors producing distortions in evaluation systems and of corresponding ramifications for students and their families.

Use of Higher Education Admissions System data in the framing of financing policy

The concept of indirect financial aid refers to a government-financing policy for institutions of higher learning based on Academic Aptitude Test (PAA) scores. Until the early 1980s, Chile's universities were more or less entirely dependent on government funding, with student registration fees and charges virtually symbolic, and revenues from sales of services and donations representing a negligible share of funding. Government financial aid was allocated based on the historical share of aid earmarked for higher education and the ability of each institution of higher learning to negotiate larger amounts of aid. This resource allocation process made for a great deal of insecurity and was responsible for the haphazard planning of educational activities by corresponding institutions.[3]

Moreover, there were no direct or indirect management audit mechanisms in place to evaluate returns on these resources, whose share of the national education budget may have shrunk relative to other budget items, but which still accounted for a large portion of spending (Lehmann, 1993). In response to the growing need for a stable financing policy and as a way of setting up certain indirect control procedures, an executive order ratified by the Chilean Congress in 1991 established the following three transfer mechanisms for government funds:

- Direct financial aid, funding technological and scientific research, and university extension programs;
- Indirect financial aid, allocated among accredited institutions of higher learning according to the number of the top twenty thousand applicants admitted to each institution;
- Refundable, long-term, low-interest student loans for students unable to pay the cost of attending private schools up front (Lehmann, 1993).

The essential resource allocation mechanism is the indirect financial aid mechanism, whose enabling legislation specifically refers to students with PAA scores ranking among the top twenty thousand who take the test (Grez et al., 1993: 1). This legislation gave private universities, previously ineligible for indirect government financial aid, access to these funds.

Until this legislation was passed, PAA scores were the only standard criterion with a comparable scale of interpretation for identifying top students, given the well-known comparability problems associated with secondary school grades, the only other possible criterion. Moreover, grades are an extremely weak basis for differentiating between students, given their highly biased nature. This latter problem stems from the fact that, knowing that grades are used as a placement criterion for admitting students to univer-

sities, teachers tend to help students as well as respond to parental pressure by giving them good grades,[4] particularly in the last years of secondary school.

This policy has affected the placement process in terms of the manipulation of weights assigned to the PAA as part of this process. In fact, the weight assigned to this test overshot scholastically recommended limits, from an average of 42.9 percent in 1979 to as high as 63.9 percent by 1982, which is more or less its current weight (Grez et al., 1995). This change in higher education policy affects the access of low-income students, because universities admit students with the best test scores on the PAA, who are mainly from middle and high-income families (Himmel and Maltes, 1981).

This use of evaluation data would appear to be symbolic or suasive . Actually, it can also be considered an instrumental use of this data. While this use of system data was not envisioned in the original design of the admissions system, it is, nevertheless, an indicator of the effectiveness of institutions of higher learning in recruiting top students and is, without question, tied to the stature of the education offered in these institutions (Himmel and Maltes, 1980). Moreover, private institutions of higher learning publish reports on the amount of aid they receive through this channel.

Establishments preparing students for entrance examinations for the higher education system

The establishment of the admissions system for higher education triggered a wave of private "preuniversity" establishments preparing students to take entrance examinations. Over the years, there has been a surge in the number of such establishments, with over half of all applicants to institutions of higher learning currently attending preuniversity courses. The nation's schools have also instituted comparable preparatory programs for placement examinations. These programs create additional expense for parents and, thus, are accessible only to students able to afford them. Nevertheless, there is no proof that attending a preuniversity course significantly improves test scores (Rojas, 1985).

Furthermore, while many of these establishments widely publicize the high test scores obtained by students enrolled in their courses, it has been found that they accept only students with very good secondary school grades. Thus, in this way, preuniversity establishments make symbolic use of system data.

Use of Academic Aptitude Test results as a requirement for applying for employment

For several years now, newspapers have been publishing job offers for positions such as mail carriers, elevator operators, drivers, messengers, and other low-level jobs requiring applicants to have taken the PAA.

Distortions in the SIMCE produced by school principals

For the past several years, in an attempt to improve their school's performance, certain principals have been discouraging students with poor grades from attending school on

days when SIMCE test instruments are scheduled to be administered This finding led to a decision not to furnish these schools with SIMCE test results (Prado, 1995).[5] Certain informal data reveal that schools resorting to this strategy reported normal attendance the following year. Nevertheless, this problem is still prevalent at the school-wide level. Thus, the current usage of SIMCE data, as well as the envisioned future use of this data, also fall under the concept of symbolic use.

Each of the examples presented above is an illustration of a symbolic and instrumental use of system data which, though followed up by action, is inconsistent with the original objectives of the corresponding evaluation system. As far as the unforeseen effects of evaluation systems are concerned, the only other point worthy of note is that they are difficult to predict, in that such systems are liable to have different effects under different sets of circumstances. However, it is likely that many of these unforeseen effects are triggered by groups affected by the system in a positive or negative manner.

Factors heightening or weakening the social impact of educational performance evaluation systems

This paper made the point that an evaluation system is likely to have a social impact when its findings are used in decision making processes that lead to measures affecting different social groups. Consequently, evaluation data will have an impact only if and when they are used, and they will have a major impact only in cases of their conceptual use. Accordingly, the following paragraphs look at different factors facilitating or hampering the use of evaluations, illustrating these factors with examples from Chile.

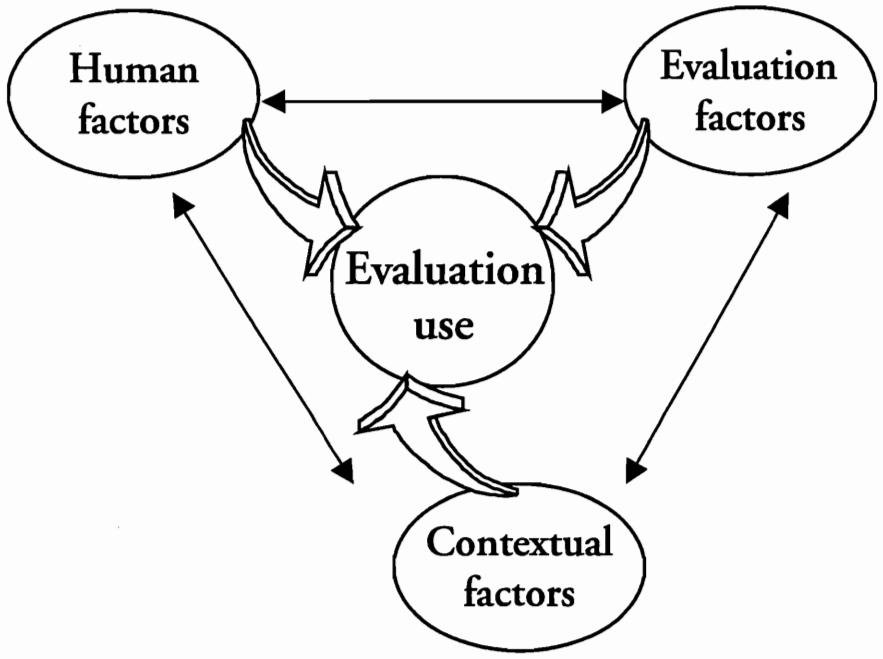
As a rule, these factors fall into one of three categories: human, contextual, and evaluation-related (Alkin, 1985). Human factors refer to the traits of evaluation system personnel and users of evaluation data. Contextual factors are stumbling blocks existing prior to the evaluation, while evaluation-related factors refer to the inherent features of the evaluation system. All these factors are interrelated, with each group of factors having specific dimensions (Figure 3).

Human factors

Chief among the traits of evaluation personnel that foster the use of evaluation data is credibility, or the confidence they inspire in system users. This trait is conditioned by their experience, their technical expertise, and their reputation within user circles. The professional team in charge of setting up the PER and SIMCE systems had credibility; its members were not on the Ministry of Education staff but on the faculty of a prestigious university. Moreover, the team in charge of developing the placement system for higher education had already distinguished itself on other assignments it had performed several years earlier.

Political sensitivity is another important trait for evaluation personnel. The team responsible for setting up the PER had limitations in this respect. In fact, it did not grasp the workings of the different political processes surrounding the implementation

Figure 3. Determining Factors in the Use of Evaluations



of the PER in time, ignored the lack of a political consensus, and relied too much on its technical expertise.

Important user traits include a commitment to the use of resulting information, in turn, conditioned, among other factors, by how necessary they consider the evaluation process to be. In the case of the PER, the teams of officials brought in by successive education ministers who had not been involved in starting up the program felt that the PER was unnecessary and, as a result, while many eventually came to accept it, they were not very committed to its use.

Another user trait that can affect the use of evaluation data has to do with professional style, which is tied in with the existence of an evaluation culture within education circles. An evaluation culture is characterized by professional educational evaluation skills, organizational leadership, and the use of resulting evaluation data in day-to-day tasks. In the case of the SIMCE and the admissions system for higher education, this evaluation culture gradually developed over time, which helped strengthen both systems.

Contextual factors

The organizational structure within which a given evaluation system is established can operate as a stumbling block in the use of evaluation data. In cases where the implementing agency is an outside organization and does not clearly pinpoint constraints imposed by the organizational structure, there can be serious opposition to the use of

resulting data. This is precisely what happened with the PER; its implementing agency failed to grasp the fact that there was an influential faction within the Ministry of Education that regarded itself as perfectly capable of conducting the PER project, and stubbornly opposed all efforts at project implementation. On the other hand, many of the members of this same faction were involved in the earliest stages of the SIMCE and were to take over the system once this phase was concluded, thereby gradually overcoming this problem.

This factor can be categorized as an intraorganizational factor. There are also extraorganizational factors that can facilitate or hamper the use of evaluation data, including the building of a political consensus around the need to implement the evaluation system. Among other things, the lack of a political consensus can threaten the system's consistent funding (Himmel, 1996). Such a consensus, which was lacking in the case of the PER, has existed for a number of years with regard to the SIMCE and the admissions system for higher education.

Evaluation factors

This category includes factors such as evaluation procedures, the stringency with which corresponding methods are followed, the features of evaluation reports, and the timeliness with which they are presented.

The procedures followed in the PER, as well as in the SIMCE and the admissions system for higher education, have all been adequate from a methodological standpoint, although none of these systems included an equivalency system for comparing consecutive assessments. In some cases, resulting reports were not sufficiently clear to their target audiences due to their use of overly technical jargon. In general, the information supplied by Chile's evaluation systems has been timely and when, for some reason, its publication was delayed, this fact was brought to light by users, some of whom maintained that the delay had made the information unusable (Zabalza, 1994).

Recommendations for national educational performance evaluation system policy

The first question is whether the evaluation system should be an integral part of the education system, which raises the second question of whether this is possible under the country's prevailing conditions.

One of the necessary conditions for the establishment of national evaluation systems is the availability of physical and human resources that can be earmarked for this task. This, in turn, requires large financial resources; thus, it necessary to ascertain whether such resources can be set aside without compromising other, possibly higher-priority investments in the education system. The availability of human resources for the implementation of a national evaluation system needs to be assured as well. In cases where necessary experts are not immediately available, the development of a team of competent professionals can take many years, and can increase the cost of setting up the system.

Another condition to considered is whether there exists a political consensus around the need to establish an evaluation system. Where it does exist, the next step is to establish the goals sought in setting up such a system. As pointed out earlier in this paper, an evaluation system can only fulfill a limited number of objectives.

Once its objectives have been defined, it is important to ascertain which social groups are likely to be affected by the system. For example, a system designed to channel students between the primary and secondary school levels into one of two tracks, (e.g., an academic track preparing them for higher education and a vocational track preparing them for the job market), would have crucial implications for students and their parents, but an indirect impact on teachers and school principals. On the other hand, a system designed to evaluate education policy would have a relatively weak impact on students and families, compared with its importance to education officials, teachers, and school principals. These examples underscore the importance of identifying the target groups affected most by the evaluation system, who would be likely to put up the strongest resistance as well as be most affected by any unforeseen effects.

Another issue that needs to be addressed is the choice of the agency to operate the system. There are a number of options, ranging from a government agency, possibly attached to the ministry of education, to a private organization. The disadvantage of entrusting this task to an agency within the ministry of education is that it appears to give the ministry a conflict of interest; it will judge the system in which it is the major stakeholder. Moreover, it is especially difficult to maintain a team of highly skilled professionals in government service in Latin American countries, with civil service pay scales so low. The result is a high turnover rate for skilled personnel, which is detrimental to a project of this sort. However, this option may have certain advantages as far as costs are concerned and in terms of heading off possible opposition within the education system.

On the other hand, a system run by an outside agency can be seriously affected by intraorganizational opposition within the ranks of the education ministry. Moreover, the evaluation system will unquestionably cost more to operate. However, it has the advantage of being independent, which enhances its credibility.

Another option is to set up an outside technical nongovernmental organization to run the system under the supervision of the ministry of education, possibly rounded out by a broad-based national committee representing different social organizations.

Assuming there is a consensus with regard to the need to set it up, and having set its objectives and identified the agency in charge, the next step is to decide the evaluation system's specific features, including but not limited to the following elements:

- *System coverage:* total population versus sampling. Should the system be administered to the entire target student population or simply to a sample of the target population? Although the objective sought by the system will be a major determining factor in this decision, it may occasionally be necessary to consider other factors as well.

Both options have their strengths and weaknesses. If a country's school population is especially large, it will be virtually impossible for it to commit the funds needed to finance such a system. However, an evaluation of the entire student population can potentially furnish each stakeholder in the education process with more relevant information, and increase the likelihood of it actually being used. Sampling, on the other hand, has the advantage of lowering costs, but the data furnished by this sort of evaluation may not live up to user expectations, which, in turn, reduces the likelihood of it effectively being used.

- *Curriculum coverage:* Any endeavor to evaluate educational achievement must address at least three issues. The first issue is to decide which grade levels should be targeted by the evaluation, since it is highly unlikely that all grade levels will be included. This necessitates dividing the school curricula into meaningful units for the formal education process. The second issue, which goes hand in hand with the first, is to decide whether the assessments per se will be limited strictly to skills, content, or objectives at the specified grade levels, or whether they will also include achievements at lower levels. Lastly, it is essential to establish which curriculum areas are to be included, as it is impossible to cover all areas of the curriculum due to time constraints, technical feasibility, and/or the cost factors.
- *Regularity with which the evaluation system is to be administered:* The effort involved in setting up an evaluation system requires that it have continuity. Thus, it is imperative that it be administered at regular intervals to furnish necessary data on a systematic basis. Systems designed to corroborate student achievement or to screen or place students will necessarily need to be administered every year. On the other hand, systems designed to evaluate education policy, provide monitoring data, or make educators accountable for student learning could be administered at wider intervals, since such changes take time and require mobilizing an enormous number of different stakeholders. Moreover, the issue of the grade levels to be targeted by the evaluation system should be considered in determining the assessment intervals, particularly in cases of systems that include monitoring. For example, assessments of students completing the third and sixth grades should be administered at intervals allowing for the same cohort to be tested at the third and sixth-grade levels.
- *Linkages with other information systems:* National evaluation systems are also information systems and are not the only such systems used in education circles. Thus, they will need to be coordinated with other educational and social information systems from the beginning to prevent overlapping in data collection efforts and facilitate the sharing of data with other systems.

This paper has examined certain problem areas associated with the social impact of national educational performance evaluation systems from the perspective of the Chilean experience. Obviously, such a study is both incomplete and biased. Nevertheless, it can still provide useful information, bearing in mind the highly-specific social effects of evaluation systems at the national level, whether expected or unforeseen, need to be examined within the specific context of an individual country.

NOTES

[1] Free private schools receive a government grant for each student, which may represent all or part of their funding.

[2] Applicants must score at least 450 points on the Academic Achievement Test to be eligible to apply to government-funded universities. However, 38 percent of the student population fails to achieve this score. Consequently, some students are reported to take the tests as many as seven times. Rounding out this group are students accepted into a program other than their first choice, who wish to transfer to another program (Rojas, 1984). Until a few years ago, there was very little flexibility for university students in Chile to transfer between different programs of study. This forced students wishing to change their field of study to retake the entrance examinations. It was even more difficult to transfer from one university to another.

[3] Until 1980, there were eight national universities in Chile, with their main campuses in one city and branches in other cities, including two public and six government-funded private universities. The reform legislation of 1980 restructured this system in such a way that all the original universities lost their branches, and eight original institutions were divided into twenty-three separate universities, fourteen public and nine government-funded private universities. All were eligible for indirect financial aid and student loan funds. Another forty-four private universities were founded over the period between 1981 and 1995. These establishments are eligible only for indirect financial aid (Muga and Rojas, 1993). Added to these universities are large numbers of public and private professional training institutes and technical training centers that have sprung up since 1981, whose stature has been rising, both in the eyes of parents and students.

[4] This phenomenon is a proven fact, with the national average of raw scores jumping from 4.7 to 5.6 on a scale of 1 to 7 over the period between 1967 and 1996, while there was virtually no change in performance on achievement tests throughout the same period.

[5] A comparison of enrollment data incorporated into SIMCE data bases three months prior to the administration of corresponding test instruments with attendance data for the scheduled testing date revealed larger-than-normal differences within a certain group of schools. Based on this data, it was decided not to furnish these schools with reports of SIMCE test results. This information was also left out of announcements published in the press.

REFERENCES

Alkin, M. (1981). "The Feasibility of Measuring Educational Attainment in Chilean Schools." Consultant Report. Santiago, Chile: PER, Report No. 8.

Alkin, M.C. (1985). *A Guide for Evaluation Decision Makers.* Beverly Hills, California: Sage Publications.

Alkin, M.C., Daillak, R.Y., and White, P. (1979). *Using Evaluations: Does Evaluation Make a Difference?* Beverly Hills, California: Sage Publications.

Braskamp, A.L. (1982). "A Definition of Use." *Studies in Educational Evaluation,* 8: 169-174.

Brown, R.D., Newman, D., and Rivers, L. (1985). "An Exploratory Study of Contextual Factors as Influences on Schoolboard Evaluation Information Needs for Decision Making." *Educational Evaluation and Policy Analysis*, 7: 437-445.

Cabezón, E., Condemarín, M., and Vacarro, L. (1996). "Jóvenes Monitores: Una Esperanza de Recuperación del Ethos Normalista." Santiago, Chile: *Revista de Educación*, 240: 50-55.

Cerda, A.M., Assaél, J., and Rodas, M.T. (1996). "Una Mirada Evaluativa a los Proyectos de Mejoramiento Educativo." Santiago, Chile: *Revista de Educación*, 236: 51-55.

Contreras, J. (1988). "Impacto del Programa de Evaluación del Rendimiento Escolar (PER) a Nivel Institucional: Establecimientos Educacionales." Masters' thesis in education science. Santiago, Chile: Pontificia Universidad Católica, Education Department.

Davis, G.A. and Thomas, M.A. (1989). *Escuelas Eficaces y Profesores Eficientes*. Madrid: La Muralla.

Enrione, A.E. (1991). "Liderazgo del Director." Master's thesis in Engineering, Pontificia Universidad Católica de Chile, School of Engineering.

Greany, V. and Kellaghan, T. (1996). *Monitoring Learning Outcomes of Education Systems*. Washington, DC: Directions in Development, The World Bank.

Grez, N., Cazenave, J., Gonzalez, M., and Gil, F.J. (1993) "Una Propuesta al Proceso de Selección de las Universidades Chilenas. Documento de Trabajo No. 7/93." Santiago, Chile: Corporación de Promoción Universitaria.

Himmel, E. (1996). "National Assessment in Chile." In Murphy, P., Greany, V., Lockhead, M., and Rojas, C. (eds). *National Assessments. Testing the System*. Washington, DC: EDI Learning Resources Series, The World Bank: 111-128.

Himmel, E., Majluf, N., and Maltes, S. (1980). "Efecto de Variables Macroestructurales y del Colegio sobre el Rendimiento en las Pruebas de Selección Universitaria." Santiago, Chile: Vice-Chancellor's Office, Pontificia Universidad Católica de Chile, Office of the Academic Vice Chancellor, Report No. 10.

Himmel, E., Majluf, N., Maltes, S., and Gutiérrez, J. (1982). "Análisis de los Efectos de Variables Socioeconómicos y del Colegio sobre el Rendimiento Escolar." Santiago, Chile: Pontificia Universidad Católica de Chile, DIUC Project No. 42/81, Final Report.

Himmel, E., Majluf, N., Maltes, S., Gazmuri, P., Contreras, J. and Escobar, C. (1988). "Análisis del Impacto sobre el Sistema Educacional de 3 Años de Aplicación del Programa de Evaluación del Rendimiento Escolar." Santiago, Chile: FONDECYT Project. Final Report.

Himmel, E. and Maltes, S. (1980). "El Prestigio de las Carreras Universitarias en Chile." Santiago, Chile: *Revista de Educación*: 33-37.

Himmel, E. and Maltes, S. (1987). "Factores que Inciden en la Decisión Profesional y el Rendimiento Académico." In *Orientación: Nuevas tendencias y perspectivas*. Santiago, Chile: Ediciones Universidad Católica de Chile: 125-151.

Himmel, E. and Maltes, S. (1981). "¿Quiénes Son los 20.000 Primeros Puntajes de la Prueba de Aptitud Académica?" Santiago, Chile: Pontificia Universidad Católica de Chile, Office of the Academic Vice Chancellor, Report No. 2.

King, J.A. and Peachman, E.M. (1984). "Pinning a Wave to the Shore: Conceptualizing Evaluation Use in School Systems." *Educational Evaluation and Policy Analysis*, 6: 241-251.

Lara, F. (1985). "Ingreso a la Educación Superior: El Sistema Nacional." Santiago, Chile: *Cuadernos del Consejo de Rectores de las Universidades Chilenas*, 24: 1-23.

Lehmann, C. (1993). "Financiamiento de la Educación Superior en Chile. Resultados del Período 1982-1992." In *Informe de la Educación Superior 1993*. Santiago, Chile: Colección Foro de la Educación Superior: 127-176.

Lockhead, M. (1996). "International Context for Assessments." In Murphy, P., Greany, V., Lockhead, M., and Rojas, C. (eds). *National Assessments. Testing the System*. Washington, DC: EDI Learning Resources Series, The World Bank: 9-19.

Lockhead, M.E. (1991). "World Bank Support for Capacity Building: The Challenge of Educational Assessment." Paper presented to the International Consultative Forum in Paris.

Ministry of Education (1997). "Resultados SIMCE 1996: ¿Qué Pasó con los Aprendizajes?" Santiago, Chile: *Revista de Educación*, 244: 20-23.

Muga, A. and Rojas, F. (1993). "Estadísticas de la Evolución del Sistema de Educación Superior en Chile: 1980-1992." In *Informe de la Educación Superior 1993*. Santiago, Chile: Colección Foro de la Educación Superior: 39-126.

Parraguez, W. (1992). "Profesores Rurales Rompen el Aislamiento." Santiago, Chile: *Revista de Educación*, 198: 16-20.

Prado, M.N. (1995). "Un SIMCE no se Puede Improvisar." Santiago, Chile: *Revista de Educación*, 198: 9-11.

Rojas, R. (1983). "Antecedentes de los Rezagados y Actividades Realizadas por Ellos Entre el Egreso de la Enseñanza Media y 1983." Santiago, Chile: University of Chile, Academic and Student Affairs Office, Monograph 19.

Rutherford, W.L. (1985). "School Principals as Effective Leaders." *Phi Delta Kappan*, 67: 31-34.

Schiefelbein, E. (1992). "Análisis del SIMCE y Sugerencias para Mejorar su Impacto en la Calidad." In Gómez, S. (ed). *La Realidad en Cifras. Estadísticas Sociales*, Santiago, Chile: FLACSO/INE/UNRISD: 241-280.

Zabalza, J., Rojas, V., and Soto, J. (1994). "Diferencias en el Rendimiento Académico en el SIMCE, en Sectores Socioeconómicos Medio-Bajos y los Principales Factores Explicativos Atribuibles al Establecimiento Escolar." Santiago, Chile: FONDECYT Project 920742. Final Report.

CHAPTER 6

THE SYSTEM FOR EVALUATING THE QUALITY OF EDUCATION IN COLOMBIA

Gabriel Restrepo[1]

This paper, like the preceding one, describes a national experience covering three decades in the application and analysis of tests of academic learning. Colombia's national testing system has demonstrated considerable stability and management capability and enjoys the confidence of both institutions and users. The tests were initially used for screening purposes and subsequently for improving the quality of education. The case study analyzes the global and national context of the transition and the mechanisms established to respond to new needs imposed by recent educational reforms

INTRODUCTION

Problems and dilemmas

Today's educational systems face a sort of paradox resulting from two apparently contradictory trends. The first is the advent of the global village (McLuhan, 1985), while the second is the increased value placed on local life. The globalization of markets and "technological convergence" (Nelson and Wright, 1992) force individuals to develop universal skills, which to a certain extent are indifferent to specific times or settings. Such skills arise out of contemporaneous scientific-technical requirements that today cancel out former comparative advantages. They make up a sort of universal citizenship, no longer induced by religion or technology but rather by instrumental rationality—a rationality that, however, is the myth *par excellence* in the contemporary world (Alexander, 1991). Such an open education is a direct result of the scope and speed of communications media and international information networks.

Moreover, the legitimacy crisis of governments and the deconstruction of the discourses of power (Weiler, 1992) call into question an education based on externally imposed rationalities. Based on an appraisal of the world of daily life (Habermas, 1987), many clamor for an education founded on local styles of knowledge and with a preference for

self-evaluation over external evaluation, for "qualitative" studies over "quantitative" studies, for affection over understanding, and for the absolute value of the individual over aggregations.

Stretched taut between universalist and local ideals, educational systems are experiencing a crisis that at certain times leads to paralysis, at others to uncertainty, and in fortunate cases to creativity. Private and public schools and universities have ceased to be self-sufficient institutions as they were prior to the scientific-technical revolution. They have instead become cogs in a more complex piece of machinery that is a replica of society itself (Coombs, 1971). Thus, the policies adopted by governments have been to simultaneously address two apparently opposing thrusts: the integration of citizens into increasingly international networks and the granting of increased autonomy to individuals and local communities.

This tension extends to national evaluation systems. Should such systems respond to universal demands or local needs? Should they limit themselves to recording knowledge and skills produced in different contexts, under the assumption that such knowledge and skills are of incommensurable worth? Since evaluation is "the power to determine the value of something" (Kvale, 1992), it ranges between the universalism of instrumental rationality and the particularisms of local preferences, with all their subjective bias. Decentralization and evaluation are desirable aspirations, but they could come into conflict if no effort is made to reconcile them (Kogan, 1992).

These tensions are experienced more dramatically in societies that have not yet accessed an economy, state, or culture in the modern sense—societies in which, given the open questioning by postmodernism, the above-mentioned crisis may easily result in paralysis, or at least perplexity.

Purpose of the essay: the case of Colombia

This essay describes Colombia's effort to construct a national system for evaluating the quality of education that, although unable to eliminate the above-described tensions, reflects both continuity over a thirty-year period as well as considerable creativity in seeking ways to overcome them.

In 1968, Colombia created a national screening system to determine access to higher education. In 1990, the country put into effect a national system for evaluating the quality of education. The transition from a screening system to an evaluation system is not easy, since they are intrinsically different (Greagney and Kellaghan, 1996; Rodríguez, 1982). Nevertheless, far from being antagonistic, they may actually be complementary. The Colombian experience provides evidence that it is possible to extract the best advantages from each.

The first part of this essay provides a summary of the evolution of education (1950-1990) in which both the limitations and opportunities for developing a national screening system to determine access to higher education are highlighted. The second

part examines the transition from a screening system to a broader evaluation system (1990-1997), within the framework of the more recent evolution of education and society. The essay then shows how both systems may actually be complementary. The third part of the essay summarizes the principal conclusions and describes policy options that may be useful to other countries, as determined by their specific characteristics.

In preparing this study, interviews were held with public authorities, directors of public and private schools and universities, professors, researchers, students, and families. A review was made of a press file containing more than eight hundred entries covering the past ten years. Existing literature on the subject was also scrutinized. The National Testing Service has provided all possible support with regard to available information.

SCREENING SYSTEM: STATE EXAMS FOR DETERMINING ELIGIBILITY FOR HIGHER EDUCATION (1964-1990)

Education in Colombia through 1990

In Colombia, "the education expectancy remained entirely unchanged at 1.4 years per person from the turn of the century through the 1950s" (National Planning Department, 1997). Up to that point, "Colombia was the country showing the greatest lag in education in Latin America" (DNP, 1991).

However, from 1950 through 1988 "Colombia experienced a growth in education coverage greater than that of any other country in Latin America during the period—including Nicaragua—and was surpassed at the international level only by the Congo, Nepal, and Togo" (DNP, 1991).

Between 1950 and 1990, primary education coverage increased from 45 percent to over 90 percent. Over the same period, coverage for secondary education grew from 5 percent to 48 percent, while that for higher education increased from 1 percent to 12 percent (Duarte, 1997). The expression of the political will to end the violence between the two traditional political parties led to a considerable increase in both public and private expenditures on education, which grew from 1 percent in 1950 to 3.5 percent in 1984 (Duarte, 1997).

In 1953, integrated educational planning was launched, and an educational credit institute for study abroad, which served to form a scientific and technical critical mass, was founded. That institute based its selection criteria exclusively on merit. In 1960, the central government assumed the costs of primary education and, in 1975, the financing of public secondary education. In both cases, it imposed controls, not always effective, on departmental and municipal governments (Duarte, 1995).

However, in 1976 the expansion came to a halt:

> Beginning in the mid-1970s, the rate of growth observed in education decreased noticeably. At the primary school level, annual

> growth rates in enrollment were less than 1 percent. In secondary school, which is where the greatest deficiencies exist, growth was scarcely 5 percent between 1975 and 1984 and 2 percent between 1985 and 1990. Education expectancy virtually ceased to grow over the last 15 years to the point that it almost reached, in the 1980s, an educational inequity greater than that found in all countries of the world with the exception of India (DNP, 1991).

The increase in the participation of education in the gross domestic product decreased beginning in 1984, and it was not for another decade that it would be possible to recover the upward trend, despite the fact that Colombia maintained a positive rate of growth during the so-called lost decade.

As a result, educational coverages are today quite precarious: 20 percent of children between the ages of 6 and 11 are not enrolled in school, while half of all children and adolescents between the ages of 12 and 17 do not attend secondary school. Of those who are enrolled, many have little likelihood of remaining in school as a result of the inefficiencies of educational institutions. The problem is more pronounced when one takes into account individual opportunities (which are fewer in number for the regions and the lower economic strata) and institutional advantages (which are fewer for rural areas, small cities, and institutions offering night courses and generally greater for private and public schools in large cities) (Duarte, 1997).

The decrease in the growth of education observed since 1976 has been the result of imperfections in the educational system, which in turn reflect weaknesses in the system of democracy. The expansion was not always the result of a rational control since, although the central government assumed responsibility for most of the costs of education, it failed to create effective instruments of control. For example, "it was not until 1991 that it had more or less reliable data to indicate how many official teachers there were and what their corresponding salary levels were" (Duarte, 1997).

Local education management was corrupted by clientelism (Gómez and Losada, 1984; Duarte, 1995). In the municipalities, individualistic transactions predominated over collective negotiations and universal norms, while a weak central government opposed a union that was strong (Duarte, 1997) and often justifiably so as a result of the precarious status of teachers.

The instability of the tenure of education ministers contributed little to rational control: the average length of service of education ministers has been one year, as contrasted with the two and one-half years for treasury ministers. This instability has been even more pronounced among departmental and municipal authorities, who rotate in response to local or regional political interests (Hansson, 1996). In addition, it is only in the last decade that economists have shown a sustained interest in problems linked to the equity and distribution of social services.

Background of the national screening system

As previously indicated, higher education coverage increased from 1 percent in 1950 to close to 12 percent in 1990, the equivalent of an additional half-million students (Duarte, 1995). In 1991, Colombia ranked at about the midpoint among Latin American countries in terms of coverage, 16.9 percent below the regional average (Alvarez, 1995). However, since the starting point was so low, annual growth rates were high. In 1989 there were a total of 236 institutions of higher learning, 30 percent of which were public and 70 percent private (DNP, 1991). In 1950, when there were no more than 40 universities, the proportion of public to private was exactly the opposite.

To regulate university expansion, the government created, in 1954, the National University Fund. In 1958, acting on their own initiative, the universities created the Colombian Association of Universities. These two institutions conducted studies on Colombian high school students, created professional orientation services, and organized four seminars on university admissions between 1960 and 1966 (Acero, 1990). In 1966, the National Testing Service (SNP) was created as an agency attached to both of the above entities. The SNP developed an ambitious strategy that included the following activities (Acero, 1990):

- Aptitude and vocational testing for the fourth year of high school
- Preparatory and skill testing for improving the quality of teaching in secondary schools
- Knowledge and skills testing for high school graduates
- Creation of admissions offices in each university
- Institution of a common basic year of university study for each of several broad areas
- Selective testing at the conclusion of each year
- Testing for university graduates

Thus, the founders had in mind a screening system that would be part of a larger quality evaluation system (Greagney and Kellaghan, 1996). The birth of the screening system was primarily the result of the evolution of psychology in Colombia, which has been exceptional (Ardila, 1973; 1993). The founders of the system were inspired by the U.S. Education Testing Service, who in 1962 received training in the construction of the Scholastic Aptitude Test (SAT) used by the College Entrance Examination Board (Caro, 1990; Rodríguez, 1982). One version of the SAT was adapted for Puerto Rico at about the same time, with assistance from Colombian psychologists. Subsequently, the tests were freely and creatively adapted to the Colombian context. Between 1964 and 1967, the Service gradually began conducting tests in a number of public universities, which served to bolster confidence in the organization. Considerable trial-and-error experience was accumulated through a process characterized by little division of labor and considerable manual activity, but no small amount of enthusiasm.

The identification of a way to finance the program, by covering program costs with inscription fees, was a landmark occurrence. As will be seen, program expansion generated its own financing.

Description of the screening system

The first national examination to screen candidates for access to higher education was conducted in 1968. In that year, the Service was attached to the Colombian Institute for the Advancement of Higher Education (ICFES), the official agency that succeeded the Fund, with administrative autonomy under the Ministry of Education's supervision.

Through 1990, the principal program conducted by the National Testing Service was the administration of examinations to determine acceptance into centers of higher learning. Other programs included the validation of high school completion or completion of specific high school grade levels for adult students or students studying under radio-based programs, examinations for acceptance into the foreign service, entrance exams for the higher levels of medicine, and issuance of certifications.

The testing process went through two phases through 1990:

1. Between 1968 and 1980, testing was voluntary for universities wishing to avail themselves of this screening instrument. However, the number of individuals taking the tests grew continuously from 26,253 in 1968 to 108,268 in 1979 (see Table 1). In 1975, 39 percent of higher education institutions requested data on applicants.

2. Beginning in 1980, the government ordered mandatory testing. However, in order to preserve university autonomy, scores could be one of a number of variously weighted criteria that each institution could take into consideration. From 1980 to 1989, coverage by the screening system almost doubled, increasing from 150,267 in 1980 to 275,152 over that period (see Table 1).

The growing demand for testing, together with the demand for other related programs, made it possible to increasingly systematize the Service, increase its technical capability and, in particular, provide multiple services that together generated a considerable surplus every year, as will be illustrated below.

The structure of the test is determined by the taxonomy of the objectives of education as established by Benjamin Bloom (Bloom et al., 1973; Acero, 1990). Of the three domains, the test examines cognitive development, disregarding affective and psychomotor development. Cognitive development is in turn divided into functions of memory, comprehension, application, analysis, synthesis, and evaluation (Bloom, 1973), which determine the structure of the questions.

The Service reflects official study programs, which through 1994 were the only ones in use in the country. The tests include a mandatory section and an elective section. The mandatory section consists of one test of verbal aptitude and another of mathematical aptitude, to which are added five tests of knowledge in biology, chemistry, physics, Spanish and literature, and the social sciences. The eighth test is an elective chosen from among twelve different options, most of which correspond to the areas stressed in the diversified high school (*bachillerato*) program (final two years).

Table 1. Number of Persons Taking the National Exam

Year	Total number of students tested
1968	26,253
1969	32,253
1970	44,339
1971	50,747
1972	51,650
1973	53,498
1974	55,662
1975	75,907
1976	80,337
1977	94,689
1978	95,757
1979	108,268
1980	150,267
1981	174,397
1982	224,335
1983	223,785
1984	223,938
1985	228,272
1986	240,442
1987	250,104
1988	265,147
1989	275,152
1990	301,073
1991	297,143
1992	306,877
1993	338,534
1994	379,827
1995	436,176
1996	480,611

Source: Historical statistical series for the programs of the National Testing Service. 1968-1997. ICFES, National Testing Service. 1990.

Through 1991, the questions were designed by psychologists, specialists in psychometrics, statisticians, and teachers from the primary, secondary and higher-level education systems in the various regions of the country.

The examination includes some 460 multiple choice questions, with one correct response and four distractive responses. Of these, there are fifty that are not taken into

account for scoring purposes, as they are included for experimental purposes for possible incorporation into future tests. There is a data base containing some thirty-five thousand questions which are being added to the computerized system.

Following the optical scanning of the answer sheets, the scoring process is done electronically (Caro, 1990). The number of correct responses constitutes the raw score, which is converted to a standard scale. The raw score for each of the eight series of questions ranges from 20 to 80, with a mean of 50 and a standard deviation of 10. In this way, a normal curve ranging between zero and somewhat more than 400 points is plotted in order to compare the differential results obtained by the examinees. Results can be grouped and compared by area, school, region, and various cohorts over time. The entire process is systematized.

Tests are administered over a one and a half day period, in three three-hour blocks, during April or May for schools on the B calendar (school year beginning in August) and during September for those on the A calendar (school year beginning in February). Students who have previously taken the test are eligible to retake it. Through 1990, tests were administered in some 148 municipalities, 488 buildings, and 8,800 testing rooms (Caro, 1990), in accordance with the up-to-date data base of high schools and institutions of higher education maintained by the Testing Service. Coordinators and panel members are chosen on a competitive basis from among teachers and university students. Testing dates are published through a number of media, including national and local newspapers. The National Testing Service includes secondary education facilities in its data base, and application forms are sent to these institutions through their directors.

The National Testing Service has access to the ICFES print shop and to modern computer systems. Results are delivered one month after administration of the test, through the school directors. Students may also request certifications by mail or by calling a central telephone number. The Service periodically publishes population statistics (traits of graduates, orientation toward higher learning), reports on the academic level of intermediate education facilities (divided into high, intermediate, and low levels of performance), and lists of students receiving the highest scores, subdivided by municipality.

The National Testing Service takes special care to preserve secrecy with regard to the design, testing, and custody of the questions developed by the teams under the coordination of the Service. The same is true for the printing, transportation, gathering, and processing of results, all of which, with the exception of the latter, are entrusted to a securities transportation firm. On only two occasions was there any risk of fraud, when test booklets were stolen, but in each case the problem was detected in time.

Impacts of the screening system

There has been a considerable degree of confidence in the results of the tests, although on occasion they have been the object of criticism, particularly with regard to their

overemphasis on the cognitive dimension vis-à-vis other components. In this regard, the Service takes great pains to insist that the tests have not been designed to evaluate the quality of intermediate education, although they may indeed be an indirect source for such evaluation (and in fact a valuable one).

It is possible to identify seven measurements of the impact of the screening tests:

1. Use of the results as a screening mechanism. Of a total 212 universities in operation in 1989, twenty-eight (13.2 percent) used the score received on the examination administered by the National Testing Service as their sole admission criterion; twenty-nine (13.7 percent) added written tests administered by the university;[2] eighty-three (39.2 percent) used the results of the examination administered by the Service in conjunction with personal interviews; fifty-six (26.4 percent) used that examination in combination with university tests and interviews; two (0.8 percent) combined the Service-administered tests with university tests and high school grades; five (2.35 percent) used the scores received on the Service tests in conjunction with interviews and high school grades; four (1.9 percent) added university tests, interviews, and high school grades to the Service-administered tests; and five (2.35 percent) complemented the examination scores with high school grades (Benavides, 1989).

2. Coverage in the mass media. Both the announcement of test dates as well as the results themselves are given ample coverage in the national and local press: after all, an effort is being made to decide the future of almost 275,000 sons and daughters in 1990, or some half million in 1997, and to indirectly grade some 5,000 secondary education facilities and, accordingly, their directors and teachers as well. The information disseminated in the press and broadcast over the radio is often enriched through debates that include politicians, technical personnel, teachers, educational directors, and parents.

3. Direct impact of the exams on families, students, and teachers. Educational facilities make every effort to obtain the highest possible average scores, as such scores are determining factors of the school's prestige, levels of enrollment and costs. Parents from middle and upper-income levels in large cities take into consideration the average scores obtained by schools. All students know that their prospects for gaining admittance to a top public or private university will depend in large measure on their score. Teachers are aware of the structure of the test.

Pressure to obtain good individual and institutional results is such that, on occasion, schools dedicate considerable time during the final year to preparing for the test. In certain extreme cases, they go so far as to expel students that might jeopardize their average.

One consequence of the importance assigned to the tests is the proliferation in the cities of nonformal educational centers that promise to successfully prepare students to take the test. In one exceptional case, such a teaching center was actually created by a university. A few, according to the interviews, offer technical orientation services. Most, however, are nothing more than commercial ploys providing memory-based rote instruction.

4. Creation of a system of rewards. The Colombian Petroleum Enterprise rewards the best high school students by financing their university studies. The Education Ministry established, in 1982, the Order of Andrés Bello, which recognizes the one hundred best high school graduates. The National University attracts the best high school students with free delivery of application forms. The best scores serve to enhance the student's résumé and are a significant criterion for the approval of educational loans (El Universal 1996). The regional and local press run features on the best high school students as well as on the best schools.

5. Use of the tests as a source of indirect information on the quality of education and on high school students themselves. State examinations are the oldest and, until 1991, almost the only source of information, albeit indirect, on the quality of education, with the exception of ethnographic studies and indicators of enrollment, retention, and grade repetition. Thus, for example, in 1984 (Rojas et al.) and in 1988 (ICFES/CENCO) two different research centers prepared, with support provided by the Service, a profile of the typical Colombian high school graduate, based on the socioeconomic information on students accompanying the application for testing.

6. Regional and municipal emulation as a result of the tests. The most significant case occurred in Antioquia and its capital city of Medellín, which have shown a marked downward trend in the quality of education resulting from the indirect measurement obtained from comparing, over time, the aggregate results of the tests (Alviar and Polanía, 1993).

Indeed, between 1981 and 1990 the national trend reflected a slight increase in the percentage of schools achieving high levels of performance (from 16.81 percent to 17.28 percent). Medellín, on the other hand, showed a noticeable drop (from 24.34 percent to 20.31 percent). Over the same period, the national trend reflected a sharp increase in the number of schools with low performance (from 38.23 percent to 46.10 percent). In Medellín, however, the trend was much more pronounced (from 36.84 percent to 57.42 percent) (Alviar, 1993).

Interpretation of results became a veritable guessing game and led to to a serious national and regional controversy (El Mundo, 1991, November 11). Eventually, the educational authorities of Medellín opted to design, for the 1996 tests, the "Medellín First in the ICFES" program. To achieve this end they organized an official massive preuniversity course, provided scholarships to the one thousand high school graduates with the best scores for study in regional public universities, and offered incentives to the best schools (Antioquia newspapers, 1996). In addition, however, using its strategic vision, the Department of Antioquia negotiated an external credit in the amount of US$40 million to improve departmental education (Marulanda, 1997).

7. Ability of the Service to resolve crisis situations. On one occasion, the Service successfully thwarted the attempt of one education minister who had indicated his desire to do away with the state test (national and regional press, 1988). Subsequently, the Service successfully countered legislative initiatives aimed at suppressing or radically

modifying the tests (House of Representatives, 1992). More recently, it was necessary to appeal to the plenary session of the Constitutional Court to annul a legal provision, sanctioned by the President, by virtue of which an additional 10 percent would be added to the scores of students performing military service (National Congress, 1993; Constitutional Court, 1996).

Net outcome in terms of successes and constraints

Two factors explain the success achieved by the National Testing Service. The first is the charismatic and technical leadership it has exercised with exceptional continuity: since its founding through 1989, the Service had only two directors. Since that time there have been four other directors who, despite not a few changes, have respected and enriched the technical patrimony of the Center.

The second factor is that the Service is self-financing. "The total real cost of the three hundred thousand six hundred thirty-seven examinations administered in 1990 was 688.5 million pesos (for that year), providing a unit cost of 2,290 pesos" (Caro, 1990). Since the rate of exchange on December 19, 1990, was 562.29 pesos per U.S. dollar, the above figures are equivalent to US$1.22 million and US$4.07 dollars, respectively. Since at that time a fee of 1,500 pesos, equivalent to US$2.66, was charged, other programs implemented by the Service were responsible for covering the subsidy. Even so, in 1991 the fee was doubled.

Through 1990, there were two primary constraints. As occurs in Latin America with evaluation systems (Horn et al., undated), the information produced has been underutilized, despite the fact that its actual use is not insignificant. This can be explained by the fact that, until 1990, there was a research deficit at both the internal and external levels of the Service.

Of greater consequence is the second constraint, which is institutional in nature and has yet to be resolved: the National Testing Service has been "tied" to the Colombian Institute for the Development of Higher Education, sometimes as a division, other times as a special unit, and still other times as a subdirectorate. Such a relationship has had negative consequences in more than one sense: It has limited the system to screening exams and similar variations thereof, while simultaneously limiting its administrative capacity to direct a national evaluation system.

On various occasions thought was given to the possibility of transforming the National Testing Service into an autonomous institution that would bring together in a single entity the functions of university screening, quality evaluation, research, and even the training of human resources in educational evaluation in general, but this concept met with resistance from the ICFES. The Service finances its operations with revenues received under its various programs and, as will be seen shortly, generates a considerable surplus, while the ICFES depends on the budget assigned to it by the government for most of its control and developmental operations.

All of this has hampered the transformation of university screening functions into quality of education evaluation functions. This shift from a screening system to an evaluation system is complex, as it involves systems that respond ideally to two different logics (Greagney and Kellaghan, 1996). Thus, based on the Colombian case, it may be said that the differences between the screening and evaluation functions are:

1. One is a system for screening and assigning future roles and rewards associated with success, while the other is a system for redistributing financial, physical, and human resources as a function of equity and achievement not yet attained, though possible. Accordingly, the former favors the so-called "Matthew effect" (Merton, 1973), i.e., the concentration of opportunities in the most favored, while the latter is designed to counter that effect. The former predicts a future behavior, while the latter anticipates an improvable behavior and shows how and where it can be improved.

2. One is guided by norms, while the other is guided by criteria. Accordingly, the former discriminates, while the second differentiates.

3. One tends to be located in the terminal phase of a prolonged cycle (thirteen years, including preschool) and in the initial phase of a second temporal sequence (ten years, if doctoral studies are included), while the other follows various cut-off points over time.

4. One tends to concentrate on contents of knowledge, abstracted from others, while the other is obliged to integrate from the outset a number of different perspectives (for example, factors associated with achievement). One tends to condense, displace, and invest (in the psychological sense) all knowledge considered valid (Díaz Barriga, 1993), whereas the other may approximate knowledge from partial dimensions, in order to once again integrate them. One tends to manipulate encyclopedic ignorance, whereas the other tends to recognize more modest dimensions, although multiple in terms of inquiries.

5. One tends to be decided from a position of authority in knowledge, whereas the other presupposes knowledge that is both ubiquitous and ongoing.

6. One receives scant feedback because its source of information, though universal (the sum total of students of the eleventh grade), is the only one available to it (the students), while the other increases feedback by multiplying evaluators (self-evaluation and external evaluation), sources, methods of inquiry (quantitative and qualitative), and users (parents, local authorities, teachers, and researchers).

7. One tends toward secrecy and reserve, while the other tends toward communication and dissemination.

This opposition may be more ductile than described, however, particularly when the two systems complement each other, as occurred beginning in 1990.

NATIONAL SYSTEM FOR QUALITY EVALUATION: 1991-1997

The framework of education: 1991-1997

Since it became self-evident in 1976, the education crisis sparked an interest in reform, particularly within the heart of both the central government and the teaching profession:

1. **Central government concern** over the fact that, although increased expenditures were increasing coverage, grade repetition and dropouts tended to cancel out achievements. To correct this situation, the government applied a number of different strategies between 1976 and 1989. Such strategies were a form of rational control—always limited—imposed by modernizing echelons possessing technical authority but little political power: the design of an educational map made up of districts and nuclei to foster supervision; curricular renovation; the introduction of experimental pilot centers; the monitoring of intergovernmental relations; the approval of the Teaching Statute and negotiation of social benefits for the teaching profession; curricular streamlining and automatic promotion; reforms made to the Ministry of Education; and the New School strategy for rural primary education.

2. **The teacher-backed pedagogical movement** determined to assimilate heterogenous sources in order to reformulate the role of both teacher and education (Caviedes, 1975; Martínez et al., 1994; Medina, 1996). No small number of studies and research efforts appealed to Marxist, structuralist, psychoanalytical, ethnographic, hermeneutic, linguistic, neopositivist, neomodernist, constructivist, and postreproductionist theories and methods, which served to enrich current thinking with regard to education (Díaz Villa, 1993). Ethnographic studies on school and youth revealed the depth of the crisis (Parra, 1996; Pérez and Mejía, 1996).

Such apparently dissimilar interests converged, however, together with many other factors, to form a new social pact that was given form in the Constitution of 1991. The latter proclaimed a social state of law with broad liberties; recognized cultural plurality; expanded civic power in the direct election of local and regional authorities and in the indirect monitoring and control of public management; and authorized the progressive decentralization of numerous basic services, provision of which was transferred to the departments and municipalities, together with the cession of an increasing proportion of national revenues.

These changes inspired the General Law of Education (Law 15 of 1994), which was drafted with the participation of the teaching profession. That law decentralized curricular responsibilities to grant autonomy to directors, docents, teachers, students, and parents in the development of institutional education projects. Intending to maintain national parameters, the law set out to strengthen the National Education Ministry, which was to establish a concerted ten-year plan, establish a national information system, identify minimum areas and indicators of achievement, and organize the evaluation system. The latter was to reconcile self-evaluation with external evaluation,

by including in the latter the evaluation of teachers, teaching methods, institutional education projects, texts, and materials.

Background of the evaluation system

Over the past fifteen years, the one factor that has had the greatest impact in terms of sparking a concern for the quality of education in Colombia and encouraging development of a system to evaluate that quality has been the New School program. This is a Colombian innovation that is international in scope, as it is a model that resolves with considerable imagination problems of quantity, quality, and efficiency. It perhaps constitutes the best alternative within the area of formal education for offering a complete, quality primary education (five years) to the poorest and most isolated population segments not only in rural areas but in the city as well.

It is impossible to summarize the program without doing injustice to its significance; the reader is referred to a minimum bibliography with regard to its evolution (Colbert et al., 1976; Torres, 1996; Schiefelbein et al., 1996; Psacharopoulos et al., 1996). This is a multigrade and modular system, textbook-intensive for students and guide-intensive for teachers, participative, inspired in the active school, and open to both the community and the ecology.

Its predecessor was the Complete Unitarian School, promoted beginning in 1960 by UNESCO's Principal Program, and derived from numerous pedagogical experiences of the twentieth century (Hernández, 1961). In Colombia, it was launched in 1962 in the small city of Pamplona. At the midpoint of the decade, a total of one hundred fifty schools existed. In 1967, the government extended the model to all one-teacher schools (Torres, 1996). The current profile of the New School was defined between 1975 and 1978, with a total of some five hundred schools operating in three departments.

The successes of the New School attracted the interest of the World Bank, which financed its expansion in two phases: the first, focusing on regional consolidation and experimentation (between 1982 and 1988); and the second, a phase of universalization (between 1989 and 1997) (Duarte, 1995). In 1985 there were some eight thousand schools, while by 1989 the number had grown to seventeen thousand nine hundred eighty-four and by 1991 to twenty thousand of the country's twenty-seven thousand rural schools (Torres, 1996).

The interest shown by the World Bank was understandable, as it had agreed to open lines of credit for education in the same year, 1979, in which the Swedish Academy awarded the Nobel Prize in economics to pioneers in the theory of human capital (Duarte, 1995).

Beginning in 1980, very high-quality private research institutes, multilateral and bilateral organizations, and the government itself alternated efforts to measure the quality of the teaching of primary education and to comprehend the significance of the New School. The SER Institute[3], with support from Canada's International Develop-

ment Research Center, designed a pioneer evaluation of achievement for third- and fifth-grade students in the areas of language and mathematics (Rodríguez, 1982). That study was the first step in the establishment of a national evaluation system. It should be noted that the researcher, who is now a consultant for a private institution, was the founder of the National Testing Service.

The research, conducted by sampling, was the first to use methodology that since then has been used for achievement analysis. At that time, the author clearly distinguished between norm-based tests, such as those conducted for the purpose of screening for access to higher education, and criterion-based tests, such as those involved in the research on achievement in language and mathematics (Rodríguez, 1982).

One of the conclusions of the research was decisive: "Generally speaking, achievement percentages in the New School, which covers basic education and essentially rural schools, are significantly higher than those for their peers in the traditional school, those in rural areas, and even those in smaller urban areas" (Rodriquez, 1982). This conclusion demonstrated the pedagogical excellence of the New School, since it operated in areas of considerable deprivation.

The following step was to prepare a methodology to explain the factors associated with achievement and specifically compare the New School and the traditional school. This was performed by the SER Institute at the request of the Ministry of Education. One of the researchers had reviewed certain assessments made by the teachers of the New School themselves and developed a certain degree of skepticism with regard to the model (Rojas and Briceño, 1982).

Nevertheless, the same researcher directed a later study, followed by a more ambitious inquiry, again at the request of the Ministry of Education (Rojas and Zoraida, 1987). Using the same criterion-based methodology established by Rodríguez for the analysis of achievement in mathematics and language, the author went even further, organizing questionnaires to be administered to teachers, directors, students, and parents with regard to associated factors such as self-esteem, creativity, civic behavior, social attitude, perceptions on marginalization, and frequency of exposure to communications media and books, among others.

The conclusions again demonstrated the New School's undeniable advantages: "at the national level, it obtained scores significantly higher than those recorded for graduates of rural schools in tests of social civic behavior, social self-concept, third-grade mathematics, and Spanish for the third and fifth grades of primary school" (Rojas and Zoraida, 1987). In addition, it was estimated that the cost was scarcely ten percent above that of the traditional school, with a greater intensity of investments in inputs such as texts, materials, and provisions.

Such evaluations must have been conclusive enough to convince the World Bank and the Ministry of Education to proceed, in 1988, with the program for universalizing primary education based on the New School model, and for the Ministry of Education

to decide to create, by means of Law 24 (approved that same year), a Division for Quality Control in Education. In 1989, this division proposed a national evaluation system based on the recommendations developed by a consultative group made up of representatives from public (Ministry of Education, National University, and Pedagogical University) and private (SER Research Institute) entities.

In 1990, another evaluation of achievement was conducted, experimental in nature and limited to three departments in which the New School program had shown notable progress. That evaluation was conducted with the participation of teachers and was administered to students in the third and fifth grades of primary school in the areas of language and mathematics.

The above-mentioned institutions did not include the National Testing Service, despite its experience with the validation of high school graduates, testing of students being promoted from fifth to sixth grade, and teacher evaluation. This omission reflected the differences and tension between a screening system and an evaluation system.

Creation and expansion of the system

In 1991, the National Testing Service was incorporated into the National System for the Evaluation of Quality, by invitation of the Ministry, thus contributing to the extension of coverage. The Service prepared achievement tests for mathematics and language at the fifth-grade level, which were administered by the SER Institute. The SER also administered tests in identical areas, designed by contract specialists, to third-grade students. Questionnaires were added to study associated factors.

The measurement of achievement was carried out in thirteen of the thirty-three territorial divisions in the country, using a sample of 15,000 students from 218 urban schools and 212 rural schools. The sample was later expanded to more children in the same grades and areas (Ministry of Education, 1993). In October 1992, tests in the same areas were applied to 25,189 students from the seventh and ninth grades in twenty-two territorial divisions, together with questionnaires on associated factors. In 1993, they were applied to 11,591 students in four departments.

The discussions that took place in November 1993 revealed that the complexity of the system resulted from the differences in viewpoint with regard to the conception and management of information, administration, evaluation, research, regional participation, dissemination, and decision making.

Between 1993 and 1994, the Ministry of Education, the National Testing Service, and the SER Institute increased the scoring scale by establishing a baseline with a national sample designed on the basis of commonly-defined criteria. This new system would serve in the future to evaluate applications in noncovered areas, such as natural and social sciences. Testing would be repeated at four-year intervals in each of the areas, in order to record changes over time.

The Service was responsible for reviewing the instruments for the mathematics and language tests and the questionnaires covering associated factors. The SER Institute applied the tests in 1993 to students from the third and fifth grades of the A calendar, while the Service applied them in 1994 to students from the B calendar. The sample comprised 53,000 students from 1,628 facilities in the thirty-three territorial divisions.

In December 1996, the need was perceived for a national master sample, since the preceding applications had each been conducted on samples of varying size and coverage. This task was entrusted to a specialized private firm using criteria defined by those institutions most involved in the system: the National Education Ministry, the National Testing Service, the National Planning Department[4], and the SER Institute.

Meanwhile, the Service, prompted by its new-found responsibility, put into effect a change. Beginning in 1992, administrative processes were standardized in a detailed manual of functions in such a way that each phase would provide precise feedback to the following phase (Páez, 1992). A significant investment in systematization occurred in 1995. While prior to that year manual activities constituted approximately 70 percent of the total, that same percentage was now applicable to computerized processes (Páez, 1997).

But these were not the only changes. Savings generated in time and resources were applied to a conceptual restructuring and to an expansion of human resources. Until 1990, the Service had had a deficit research capability. Subsequent to that date, permanent lines of research were created that brought together qualitative and quantitative concepts. Up until 1991, the Service was dependent on psychology. Without detriment to the crucial role of psychology, specialized teams were created in the four fundamental areas of language, mathematics, science, and social sciences. Interdisciplinary work was quite intense. The discussion was launched with teachers, universities, and regions.

The Service efficiently and creatively undertook the new challenges of the national evaluation system, without neglecting its responsibilities involving the screening system. Since 1995, Bloom's taxonomy has been reconciled with the introduction of four skills deriving from the theory of communicative action (Habermas, 1987): communicative skills, scientific-technical skills, esthetic skills, and ethical skills (ICFES. SNP. 1997, a, b, c, d, e). Beginning in 1994, the Service took into consideration, for purposes of administering the traditional exam, the minimum contents and indicators of achievement established as a national standard by the Education Ministry. In 1990, a total of 301,073 students presented for the admissions exam, a figure that ballooned to 480,611 in 1996 (see Table 1).

The increase in the coverage of this program has led to a structure that ensures the financial solvency of the National Testing Service. In 1996, the Service recorded a surplus of 1,572 billion pesos, or almost US$1.5 million (see Table 2).

The Ministry of Education and the National Testing Service assumed an even greater challenge: participation in the Third International Mathematics and Science Study

(TIMMS). The results (between fortieth and forty-first place), though predictable given the existing problems in terms of the quality of education, were not as important as the considerable experience gained with regard to the international comparison of Colombia's curricula (which are 80 percent consistent with international patterns), the increase in the technical and administrative capacity of the evaluation system, and the availability of a measure of international, national, and regional comparison to be added to existing measures. Studying and interpreting the data would provide an opportunity to introduce improvements to education (Díaz, 1996).

For the same reasons in 1995, Colombia joined the group of countries conducting an international study on civic education, with guidance provided by the International Association for the Evaluation of Educational Achievement. This involvement coincides with a research effort sponsored by the Testing Service with regard to the teaching of the social sciences, which will use both quantitative and qualitative methods. In this way, the Service is planning to extend achievement testing to domains other than mathematics and language. During the current year, the Ministry of Education and the Service will also conduct evaluations of primary school teachers. Additionally, testing in the natural and social sciences is being planned for the following year.

Impact

The greatest impact of the national evaluation system involves anticipating and justifying the decision to extend the New School program to the entire country. Coverage expanded from eight thouand schools in 1985 to twenty thousand in 1991. In addition, however, the national evaluation system has generated a methodology for assessing differences in achievement and associated factors. Such a methodology could be extended to the study of various pedagogical strategies, both public and private, as well as to secondary education.

The above has led to extraordinary progress in terms of social indicators and the theories regarding the impact of public expenditures in Colombia in general and regarding the importance of expenditures in education in particular. A joint study conducted by the World Bank and the Social Mission of the Government of Colombia led to the preparation of a very precise forecast of the magnitude of poverty (May, 1996). Certain parallel studies (Londoño, 1995; Vélez, 1996) have examined in detail the relationship between the distribution of income, public expenditures, and economic and social development. Government planners, researchers from macroeconomic and social institutions, and the benefactors of social foundations use indicators of quality with increasing frequency. One conclusion becomes increasingly certain: a well-oriented investment in education contributes simultaneously to overcoming poverty and to generating economic growth. This is the conclusion reached by the various publications of the journal *Coyuntura Social,* a biannual periodical published since 1989 by Colombia's two most prestigious economic and social research institutions (FEDESARROLLO[5] and the SER Research Institute), and inspired in large measure by the progress achieved with the evaluation system.

Table 2. National Testing Service: Program Revenues and Costs in 1995 U.S. Dollars

Program	Users	Fee	Revenue	Cost
Government Examinations CY 951	101,315	6.5	658,548	713,082
Government Examinations CY 952				
Individuals	63,316	7.1	449,544	
Schools	271,296	7.1	1,926,202	2,432,978
High School Academic Validation	62,754	39.3	2,466,232	569,208
Intermediate School Validation	1,611	30.5	49,136	15,828
Basic Education Validation	2,808	26.1	73,289	26,110
Academic Validation of Individual High School Grades	20,200	7.1	143,420	185,615
Validation of Primary Basic Education	2,556	3.3	8,435	16,037
Examinations for Entrance to the School of Medicine	1,400	9.4	13,160	8,740
Examinations for Entrance into the Foreign Service*	91	18.8	1,711	4,346
Issuance of High School Diplomas	10,626	9.1	96,697	**
Issuance of Certifications and Transcripts of Grades	77,998	1.4	109,197	**
Issuance of New Certifications	129	17.3	2,232	**
Issuance of Transcripts of Grades	108	1.4	151	**
TOTAL	616,208		5,997,952	

*The deficit in this application was covered by the Ministry of Foreign Affairs in 1996. The rate of exchange for U.S. dollars was 1,000 pesos per dollar.
**Costs included within those of each program.

Constraints

Why, then, have the results of the national system for evaluating the quality of education not been used for policy and social programming? Or, to express it another way, why, despite the advances recorded in the national evaluation system, is its use not restricted even more? There is, in the first place, a problem of coordination—very obvious to be sure—between social and macroeconomic policy and the national evaluation system. The National Evaluation System is governed by a hybrid: the Ministry of Education establishes guidelines and makes available specific resources for evaluation (which are different from those for the screening system), but technical and operating capacity rests with the National Testing Service, which is subordinated to a third institution, the ICFES (for which the activities of the Service are not particularly appreciated, even though its financial sufficiency may be envied). Nevertheless, a solution to this difference is now discernible, as the incumbent Minister of Education has established the objective of creating an institution endowed with financial, technical, and operational autonomy to measure the quality of education.

Even so, a second, more significant problem, which will require a greater degree of policy and technical dedication, is the way that the decentralization of social services, and particularly of education, was carried out. The principal defect is that the territorial management of the resources transferred is not organically linked to the management of education, since the law did not provide public or private schools with financial or administrative instruments to exercise any degree of autonomy (Duarte, 1997; Sarmiento and Vargas, 1997). In other words, unlike what has occurred in other countries, decentralization has not been designed to mobilize financial and human resources as a function of an optimal combination of needs and quality of service. This has led to an aberrant situation by virtue of which "the poorer the municipality, the fewer teachers are assigned to it by the department" (Sarmiento and Vargas, 1997), thus limiting the power of a national system for evaluating the quality of education as an indicator of the distribution of resources.

CONCLUSIONS

This study has highlighted the difference that, according to international experts, exists between a national screening system and a national evaluation system. It is possible for one system to give way to the other, but since different logics are involved, it would perhaps be more appropriate to establish from the outset an evaluation system of which screening would be but one element.

Nevertheless, the relationship between one and the other is dependent on historical and institutional contexts. In the case of Colombia, the principal advantage of having begun with the screening system is that it provided the country with invaluable experience in terms of technical and organizational skills, including the administration of a highly efficient computerized system. This system achieved, in 1996, a considerable impact, in view of the fact that it applied tests of knowledge to determine eligibility for access to higher education to almost a half-million students.

A second advantage, no less important, lies in the fact that the screening system has managed to solidify its financial self-sufficiency, which will always be a requirement for strategic continuity, and to attain a certain degree of technical independence. Given the existence of a considerable elasticity as determined by the potential growth of higher education, and given that the costs of the various screening exams may be paid for by a population for which access to higher education requires an extraordinary subsidy, many countries of the region might opt for this approach, taking care to gradually develop a parallel capacity to evaluate primary and secondary education. The key element of this point, however, lies in selecting the most relevant problem, as was the case with Colombia's New School.

A decisive matter in any case will be the institutional definition of the National System for the Evaluation of Quality. An optimal solution would be to achieve what perhaps may soon be a goal of Colombia: to create an independent institute—preferably mixed—endowed with the greatest possible degree of technical skill and social credibility. In the case of Colombia, the partnership established some time ago with the SER Research Institute proved to be most encouraging.

Finally, it behooves us to stress the importance of the capacity for dialogue among social policies, in particular education and macro-economic policies. The obligatory means for such dialogue are information and evaluation systems. Policies that are well-designed in this regard and professionally and technically competent may constitute a more effective transactional approach than endless speeches on the need for social expenditure.

Such a dialogue depends to a large extent on the way in which a country's decentralization has been conceived, and in particular on the way in which the educational equation distributes responsibilities among the central government, the departments, and the municipalities. There is in this a political dimension that transcends the technical dimension, but there can be no doubt that the latter, provided that it is properly oriented, can influence political change.

NOTES

[1] Professor at the National University, Bogotá, Colombia. The author wishes to extend special thanks to the following individuals who, among many others, kindly contributed to this essay: Benjamín Álvarez (AED); Magdalena Mantilla (Chief of the National Testing Service Subdirectorate of the Colombian Institute for the Development of Higher Education); Fernando Páez, of the National Testing Service (SNP), a subdirectorate of the Colombian Institute for the Development of Higher Education (ICFES), which is turn attached to the Colombian Ministry of National Education; officers of the National Testing Service; Blanca Otálora (Director of School Organization of the Ministry of National Education); Jesús Duarte (Chief of the Social Development Unit of the National Planning Department); Alfredo Sarmiento (Chief of the Social Mission of the National Planning Department); Pedro Amaya, Director of the SER Research Institute; Carmenza Bulla, for conducting interviews; educational directors; and university professors. However, the statements made in this document are solely those of the author and do not necessarily reflect the views of the institutions or individuals consulted.

[2] The National University used its own admissions exam, which was similar to the one administered by the Service.

[3] SER Research Institute, a private, independent, not-for-profit research entity created in 1974 for the purpose of studying institutional systems and, in particular, for conducting pioneering research on topics involving justice, civic security, social security, education, health, transportation, and public management. It has on file more than 300 reports on strategic sectors.

[4] National Planning Department (DNP), which dates back to 1950. At present, it is an administrative department attached to the Office of the President of the Republic that performs the function of secretariat for the National Council on Economic and Social Policy, headed by the President and responsible for defining investment plans, subject to a general development plan that must be approved by the Congress and subsequently subjected to a process of national consensus-building.

[5] FEDESARROLLO, a private, independent, and not-for-profit research entity, created in 1970 and specializing in macroeconomic policy. In addition to many economic reports and books, it publishes the journal *Coyuntura Económica* and, since 1989, in collaboration with the SER Research Institute, the journal *Coyuntura Social.* Its history has been recounted in the excellent book by Gómez Buendía, Hernando (ed): Economía y Opinión. Bogotá: Tercer Mundo.

[6] ICETEX. Colombian Institute for Technical Studies Abroad, an entity attached to the National Ministry of Education. It was created in 1950 and began operating in 1953. Its founder, Gabriel Betancur Mejía, designed this innovative institution based on his own reflections as a student in the United States, in about the year 1944, as the recipient of a special loan from an Antioquia corporation.

REFERENCES

Acero, Hugo. (1990). "Desarrollo Histórico de las Pruebas de Admisión en la Universidad Colombiana. Servicio Nacional de Pruebas." Bogotá, mimeographed (ICFES, SNP).

Alexander, Jeffrey (1991). (1988). "Sociología Cultural: lo Sagrado y lo Profano en el Discurso Tecnológico." In *Revista Mexicana de Sociología* (México), April-June: 283-309.

Álvarez, Benjamín. (1995). "Equity and Selective Access to Higher Education in Latin America." Kellaghan, Thomas (ed). (1995). *Admission to Higher Education: Issues and Practice.* Dublin, Educational Research Centre. Princeton, NJ: International Association for Educational Assessment: 210-222.

Alviar, Mauricio and Polanía, Doris. (1993). "La Calidad de la Educación." In *Educación y Cultura* (Bogotá), 29, 24-36.

Ardila, Rubén. (1973). *La Psicología en Colombia. Desarrollo Histórico.* México: Trillas.

Ardila, Rubén. (1993). *La Psicología en Colombia. Contexto Social e Histórico.* Bogotá: Tercer Mundo.

Benavides, Alvaro. (1989). *Admisión a la Educación Superior: Panorama Internacional.* Bogotá: Universidad Nacional.

Betancur Mejía, Gabriel. (1997). "Entrevista Concedida a Gabriel Restrepo Sobre el Planeamiento Integral de la Educación, la Creación de ICETEX[6] y la Evolución de la Educación." Bogotá, typewritten.

Bloom, Benjamin et al. (1973). *Taxonomía de los Objetivos de la Educación. La Clasificación de las Metas Educacionales.* Buenos Aires: El Ateneo Editorial.

Caro, Blanca Lilia. (1990). "Impacto del Servicio Nacional de Pruebas." Bogotá: Instituto SER, IFT-196, typewritten.

Caviedes, Sergio. (1975). *Tecnología Educativa y Satélite Educativo.* Bogotá: Sudamericana.

Colbert de Arboleda, Clara Victoria; Mogollón, Oscar; and Levinger, Beryl. (1976). "Hacia la Escuela Nueva: Unidades de Instrucción y Formación para el Maestro Rural que Tiene Más de un Nivel a Su Cargo." Bogotá: Ministerio de Educación Nacional y Universidad Pedagógica, typewritten.

Congreso de la República. (1993). "Ley 48 por la Cual se Reglamenta el Servicio de reclutamiento y movilización." Bogotá, typewritten.

Coombs, Philip. (1971). *La Crisis de la Educación Mundial.* Barcelona: Península.

Corte Constitucional. (1996). "Demanda Inconstitucional Contra el Artículo 40, Literal b, de la Ley 48 de 1993." Bogotá, typewritten.

Departamento Nacional de Planeación. Misión Social. (1997). "Seminario Logro del Plantel. Notas Sueltas." Bogotá, typewritten.

Díaz, Carlos Jairo. (1996). "TIMSS. Informe Ejecutivo." Cali, typewritten.

Díaz Barriga, Angel (ed). (1993). *El Examen: Textos para Su Historia y Debate.* México: Universidad Nacional Autónoma de México.

Díaz Villa, Mario. (1993). *El Campo Intelectual de la Educación en Colombia.* Cali: Universidad del Valle.

Duarte, Jesús Hernando. (1995). "State Education and Clientelism in Colombia (The Politics of State Educational Administration and of Implementation of Educational Investment Projects in Two Colombian Regions)." Thesis submitted in partial fulfillment of the requirements for the degree of D. Phil. in Politics in the Faculty of Social Studies at the University of Oxford Trinity Term.

Duarte, Jesús. (1997a). "Entrevista Concedida a Gabriel Restrepo sobre el Sistema Nacional de Evaluación." Bogotá, magnetic tape.

Duarte, Jesús. (1997b). "Situaciones Críticas del Sector Educativo Colombiano y Retos Hacia el Futuro." Bogotá, typewritten.

Gómez Buendía, Hernando, and Losada Lora, Rodrigo. (1984). *Organización y Conflicto: la Educación Primaria Oficial en Colombia.* Ottawa: CIID.

Greagney, Vincent, and Kellaghan, Thomas. (1996). *Monitoring the Learning Outcomes of Educational Systems.* Washington, DC: The World Bank.

Habermas, Jürgen. (1987). *Teoría de la Acción Comunicativa.* Madrid: Taurus. Two volumes.

Horn, Robin; Wolff, Laurence; and Vélez, Eduardo. (Undated). *Developing Educational Assessment Systems in Latin America. A Review of Issues and Recent Experience.* Washington, DC: World Bank.

Kogan, Maurice. (1992). "El Ajuste de la Evaluación en el Marco de la Dirección de la Educación." In *Revista de Educación* (Madrid), 299: 155-168.

Kvale, Steinar. (1992). "La Evaluación y la Descentralización de los Conocimientos." In *Revista de Educación* (Madrid), 299: 119-141.

Londoño de la Cuesta, Juan Luis. (1995). *Distribución del Ingreso y Desarrollo Económico.* Bogotá: Tercer Mundo.

Martínez B., Alberto; Noguera, Carlos; and Castro, Jorge O. (1994). *Currículo y Modernización. Cuatro Décadas de Educación en Colombia.* Bogotá: Foro Nacional por Colombia-Tercer Milenio.

Marulanda, Iván. (1997). "Mataron un Niño en el Japón y Dos Mellizos en Colombia." *El Espectador* (Bogotá), July 5, 1997: 2A.

May, Ernesto (general coordinator). (1996). *La Pobreza en Colombia. Un Estudio del Banco Mundial.* Bogotá: Tercer Mundo.

McLuhan, Marshall (1985). "La Galaxia de Gutemburg." Génesis del Homo typographicus. Barcelona: Planeta.

Medina Gallego, Carlos. (1996). *Caja de Herramientas para Transformar la Escuela.* Santafé de Bogotá: Rodríguez Quito Editores.

Merton, Robert K. (1973). "El Efecto Mateo." In *La Sociología de la Ciencia. Investigaciones Teóricas y Empíricas.* Madrid: Alianza. Two volumes.

Nelson, Richard R., and Wright, Gavin. (1992). "The Rise and Fall of American Technological Leadership: The Postwar Era in Historical Perspective." In *Journal of Economic Literature*, 30: 1931-1964.

Páez, Fernando. (1992). "Manual de Procedimiento para la Organización y Administración de Exámenes en el Servicio Nacional de Pruebas." Bogotá, graduate thesis submitted to the Escuela Superior de Administración Pública, typewritten.

Páez, Fernando. (1997). "Entrevista Sobre la Administración del Servicio Nacional de Pruebas." Bogotá, magnetic tape.

Parra, Rodrigo. (1996). *Escuela y Modernidad en Colombia.* Bogotá: Tercer Mundo. Four volumes.

Pérez, Diego, and Mejía, Marco Raúl. (1996). *De Calles, Parches, Galladas y Escuelas. Transformaciones en los Procesos de Socialización de los Jóvenes de Hoy.* Bogotá: CINEP.

Psacharopoulos, George, Rojas, Carlos and Vélez, Eduardo. (1996). "Evaluación de Resultados en la Escuela Nueva de Colombia. ¿Es el Multígrado la Respuesta?" In *Revista Colombiana de Educación* (Bogotá, Universidad Pedagógica), 32: 93-110.

Rodríguez, José. (1982). "El Uso de los Exámenes del Servicio Nacional de Pruebas Como Indicadores de la Calidad de la Educación Secundaria." Bogotá: Instituto SER de Investigación, IFT-040, typewritten.

Rojas, Carlos and Briceño, Rosa Cecilia. (1982). "Escuela Nueva 1977-1981. Características Generales." Bogotá: Instituto SER de Investigación, IFT-039, typewritten.

Rojas, Carlos and Castillo, Zoraida. (1987). "Evaluación del Programa Escuela Nueva en Colombia. Características de los Planteles Docentes y Allumnos; Logro de sus Estudiantes en Comparación con los de las Escuelas Tradicionales." Bogotá: Instituto SER de Investigación, IFT-133, typewritten.

Rojas, Carlos, Amézquita de Pardo, Helena and González Peña, Germán. (1984). "Características del Bachiller Colombiano y su Relación con los Resultados en los Exámenes de Estado y su Ingreso a las Instituciones de Educación Superior." Bogotá: Instituto SER de Investigación, IFT-068, typewritten.

Sarmiento, Alfredo and Vargas, Jorge Enrique. (1997). "Descentralización de los Servicios de Educación y Salud en Colombia: Versión Preliminar." Bogotá, typewritten.

Schiefelbein, Ernesto, Vera, Rodrigo, Aranda, Humberto, Vargas, Zoila and Corco, Víctor. (1996). "En Busca de la Escuela del Siglo XXI: ¿Puede Darnos la Pista la Escuela Nueva en Colombia?" In *Revista Colombiana de Educación* (Bogotá: Universidad Pedagógica), 32:20-91.

Torres, Rosa María. (1996). "Alternativas Dentro de la Educación Formal: el Programa Escuela Nueva en Colombia." In *Revista Colombiana de Educación* (Bogotá: Universidad Pedagógica), 32: 1-19.

Weiler, Hans. (1992). "¿Es la Descentralización de la Dirección Educativa un Ejercicio Contradictorio?" In *Revista de Educación* (Madrid), 299: 57-80.

Section III
TEACHER EVALUATION AND PROFESSIONALISM

CHAPTER 7

THE EVALUATION OF TEACHERS

Carol Anne Dwyer

The third part of this book begins with a critical review of the subject of teacher evaluation within the general framework of educational improvement, which is the ultimate purpose of evaluation. This chapter analyzes the purposes of teacher evaluation, the standards to which it is to be held, and available methodologies. It also identifies issues to consider when implementing teacher evaluations, such as choice of method, inclusion of research results, different emphases between theory and practice, the involvement of various perspectives, and the scope of the evaluation criteria.

INTRODUCTION

This paper discusses the following aspects of teacher evaluation:[1]

- Using teacher evaluation for educational improvement in an integrated evaluation system
- Relationships among goals, standards, and assessment in teacher evaluation
- Standards for validity, technical quality, and use; curriculum; teaching knowledge and skill; level of performance; and opportunity to learn
- Purposes of teacher evaluation
- Criticisms of teacher evaluation alternatives
- Analysis of the domain of teaching for evaluation purposes
- Role of a guiding concept of teaching for teacher evaluation
- Selecting assessment methods
- Selected issues and debates in teacher evaluation

TEACHER EVALUATIONS AND EDUCATIONAL IMPROVEMENT

In order to be effective in improving education, teacher evaluations must be treated as an intrinsic element of the educational system. This is a general principle of effective evaluation and a major factor in determining the validity of the assessments used. Evaluations of teachers have links to important social values held by the general public,

those who govern, and professional educators. Teacher evaluations are also linked to the policies and practices of teacher education; to developing competence in practicing teachers; and to decisions about school curricula. Making these links clear and coherent is the *sine qua non* of teacher and student evaluations in terms of validity, fairness, practicality, and utility.

In addition to the technical and ethical needs served by a unified vision of education, there is a practical need as well. Based on many years as a designer and developer of educational evaluations, I have concluded that data from evaluations that are treated as separate from the values held by society, and the life of schools, and the communities in which they function, will seldom be perceived as useful by their intended beneficiaries. To benefit the educational system as a whole, assessments must be carefully linked to these larger entities. The technical characteristics of teacher and other evaluations do not, by themselves, tell us whether the evaluations will be worthwhile. The value of information derived from evaluations will ultimately be determined by the extent to which the educational system is influenced—either positively or negatively.

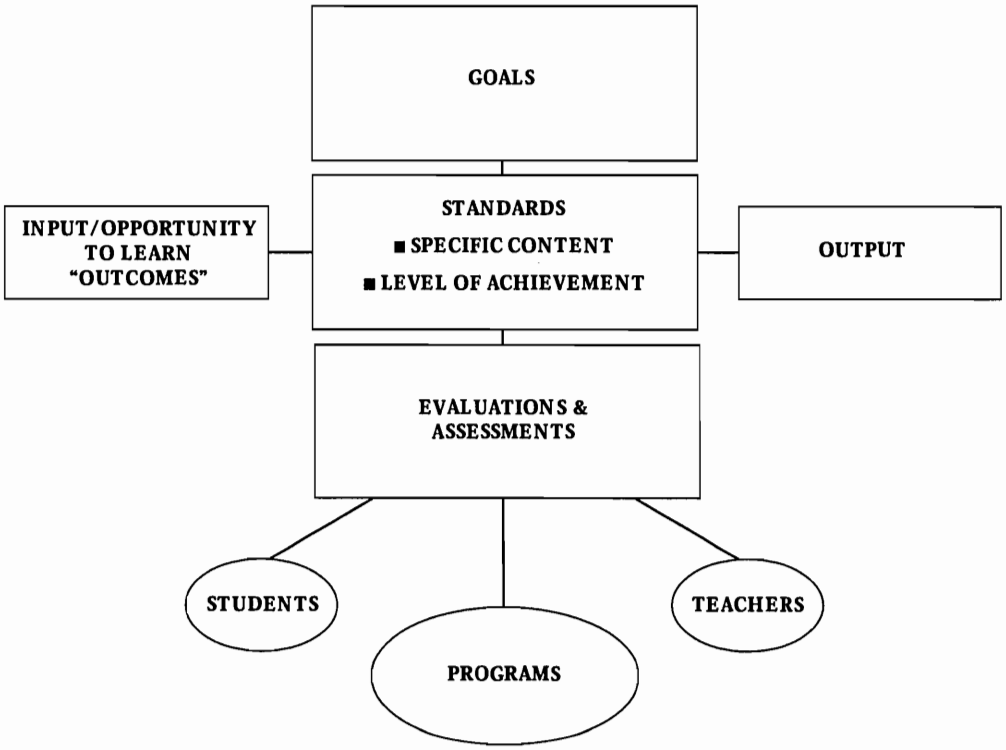
GOALS, STANDARDS, AND ASSESSMENTS

Understanding the role of teacher evaluations in educational improvements requires understanding of the relationships and distinctions that exist in the educational system as a whole. Integral to this understanding are the following three elements: goals, standards, and assessments, all of which are important contributors to educational improvement. Although some differences in their purpose, appearance, and use may seem obvious, others are not. Establishing logical links among these elements, and developing policies and practices consistent with this larger view, are critical but often neglected steps in establishing effective mechanisms for the planning, delivery, and evaluation of instruction. Figure 1 gives a simplified view of the interconnections of goals, standards, and assessments in educational improvement.

Goals are inherently aspirational statements. As such, they have three characteristics that are important for this discussion: 1) relative to standards and assessments, agreement on goals is easy to obtain; 2) goals will not always be completely achieved; and 3) there will always be important educational goals, or aspects of particular goals that, for a variety of practical or conceptual reasons, will not be measured. No matter how carefully goals are specified, they are not by nature standards, nor are they assessments. The process of moving from goals to specific assessment activities is typically complex, encompassing the many issues of resources and implementation. Thus, in the final analysis, assessments will never represent the desired goals with complete coverage or absolute fidelity. Nevertheless, the process of setting goals is a valuable endeavor, as is a well-designed assessment process.

Although many important goals prove not to be directly measurable, the existence of specific goals serves a useful function in coordinating efforts and planning the use of indirect and complementary sources of data to track progress toward their attainment. Explicit goals are also of use in resolving differences about specific policies and prac-

Figure 1. Process of Utilizing Evaluation for Educational Improvement



tices. For example, much of the effort now being devoted to improving the teaching of mathematics and science in the United States stems from international comparisons of students' achievements in these areas that reflected unfavorably on the United States. These data enabled educators in the United States to propose a goal for overall improvement that was acceptable to the general public, and, over time, to marshall considerable resources for the long process of setting the standards, creating the assessments, and providing the resources that will be needed to achieve this goal.

Once goals have been agreed upon, specific standards must be set for achieving these goals. To return to the previous example, progress toward a national goal concerned with improving students' comprehension of mathematics and science requires setting curriculum standards in both disciplines that specify what students should learn, what teachers should teach, and the level at which students should demonstrate that achievement. In the United States today, there are at least two perspectives on setting standards: the perspectives of the individual states that will implement the standards, and the perspectives of education professionals, such as the national subject-matter associations, whose teacher, teacher education, and researcher members have appropriate curricular expertise. Standards for teachers must also be developed and coordinated with the standards for student outcomes, since it is reasonable to assume that teachers cannot effectively teach what they do not themselves know. Outcome standards for teachers

may have implications for the content of their teacher education programs. Teacher content standards are concerned with the subject matter and pedagogical knowledge teachers should possess, and their ability to transfer this knowledge to students in the classroom. Input, or opportunity to learn standards, ensures that teachers have access to the continuing educational experiences, materials, and other resources that they and their students need to produce the required educational outcomes.

In standards currently being set in the United States, several important issues have emerged that are likely to be applicable elsewhere. The process of setting standards not only clarifies what is to be taught and learned, but it also raises the level of expectations about performance. Standards set in a consensual process that includes education professionals tend to be sophisticated, demanding, and forward-looking. Although the creation of high standards is a highly valued outcome of the standards development process in terms of creating a climate of high aspirations for student learning, it also creates a number of practical problems. Teachers, teacher professional organizations, and teacher educators will be justifiably concerned about how they will meet the high standards. As a practical matter, they will be concerned about how to obtain the resources that may be required to do so.

Nevertheless, these concerns have not diminished their interest in developing and implementing standards. Specifically, many teachers are not currently prepared to teach at the level that the standards imply, so extensive in-service education for teachers (and other educators in their teacher education programs) will be needed to implement the standards. In addition, if the same high level of performance is to be expected from teachers in a wide range of schools and students, attention must be given to disparities in resources (Banks, 1997). This means that standards must be concerned not only with the end-product of student learning (educational output), but also with the resources available to students and teachers to aid in teaching and learning (educational input or opportunity-to-learn standards).

This attention to input may seem self-evident in theory (if people are going to be judged on what they have learned, they should all have the same opportunity to learn it), but the implementation is highly controversial. Areas of disagreement that have surfaced in this regard include who should have responsibility for ensuring equality of input, and how funding should be provided. In addition, methodological debates have proliferated around the topic of opportunity to learn. Some aspects of input standards are relatively easy to measure, such as availability of books and the quality of the school facilities. Others are much harder to measure, however, such as teacher quality. In the United States, development and implementation of input standards has been highly controversial, even in the planning stages, and promises to continue to be so, given the wide variance in how these standards are being conceptualized.

Because assessments flow directly from standards, in principle, they should be created only after goals and standards have been specified. In practice, however, the process tends to be iterative. For example, often the first time that the aspirations of educational goals are operationalized is when they are translated into assessment tasks or test

questions. Although this is unfortunate in some respects, this situation can have positive aspects if a certain amount of flexibility exists in the educational system. Experience with writing and trying out assessments can provide insight into how standards should be refined. When standards are operationalized into specific tasks and questions for inclusion in a test, then people criticize them as too hard, too easy, irrelevant, unfair, and politically incorrect. This criticism, however, can provide an opportunity to sharpen and focus the meaning of goals, standards, assessments, and their interrelationships.

The linkages among all of the elements that contribute to achieving educational goals—students, teachers, and curricula—must be established in order for teacher assessments to be most effective. It is important that these links be made at the level of goals and standards, not just at the assessment level. At this stage of standards development in the United States, despite the separate issue of difficulties created by the many differences in format, language, level of specificity, and intent, numerous instances of broader substantive discrepancies are being identified that will need to be dealt with: discrepancies among professional groups', states' and national standards; discrepancies between what students are being asked to learn and what teachers are being asked to prepare; and, even more important, discrepancies between present and future requirements for teachers and students.

STANDARDS

There are numerous types of standards that are applicable to teacher evaluation, and there is much confusion generated by the lack of attention to the different purposes they serve. Examples of various types of standards follow.

Standards for validity, technical quality, and use

The quality of teacher evaluations can be judged by the technical standards that are broadly applied to any educational or psychological assessment. Modern validity theory (e.g., APA, 1985; Cole and Moss, 1989; Messick, 1989; Moss, 1992) emphasizes the broad context in which evaluations are developed and used, and provides conceptual guidance in considering the validity of assessments of complex activities such as teaching. This view of assessment quality highlights the importance of demonstrating logical linkages among parts of the educational system, and a systematic examination of the consequences of carrying out assessments. This view is now widely accepted by measurement specialists. The most widely referenced set of standards for assessments, *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council for Measurement in Education, 1985), reflects this point of view, as does the revision of these standards that is now underway (expected completion date 1998). These standards are, however, at a level of generality that leaves much to the discretion of those who develop and use teacher assessments. Additional literature on technical standards and expectations in this area includes that of the Joint Committee on Standards for Educational Evaluation's *Personnel Evaluation Standards* (1988). These standards pay attention not only to

assessment quality, but also to educational context characteristics and ethical use of the assessments.

The National Board of Professional Teaching Standards (NBPTS) has also offered a perspective on technical and other standards specific to the assessment of teaching that has been influential in recent discussions of teacher evaluations. The NBPTS has applied to their own efforts the standard that: "The assessments must be professionally credible, publicly acceptable, legally defensible, administratively feasible, and economically affordable" (NBPTS, 1991, pg. 53).

Proposed standards relating specifically to the technical quality of performance assessments are increasingly becoming available, although none have yet reached the level of general acceptance of the AERA/APA/NCME or the Joint Committee's standards. Proposed standards related to various aspects of teacher performance evaluation have been offered by Claxton, Murrell, and Porter (1987); Linn, Baker, and Dunbar (1991); Moss (1992); Miller and Legg (1993); and Quellmalz (1991). Dwyer (1994) offers a further analysis of this literature and its application to teacher evaluations.

Curricular standards

Curricular standards refer to sets of statements of knowledge and skills that are to be learned by students, ordinarily presented in a framework indicating the scope of the content to be covered and the sequence in which the material is generally to be taught. Curricular standards are relevant to teacher evaluations in two ways. First, sets of student curricular standards, such as those developed by the National Council of Teachers of Mathematics (1989), in defining what students are to learn, have direct implications for what teachers must know and be able to teach. By implication, such student-oriented curricular standards also pertain to teacher professional development activities that are needed. Second, curricular standards can also be developed for teacher education. Such standards refer directly to content and content-specific pedagogical knowledge and skills that should be possessed by teachers. To be effective, such standards should link these to the knowledge and skills that are to be developed in students.

Standards for teacher knowledge and skills

As noted above, curricular standards may directly or indirectly specify knowledge and skills to be required of teachers. Standards for teachers' knowledge and skills are conceptually linked to these curriculum standards, but approach the issue in a way that 1) relates directly to teachers, and 2) specifies knowledge and skills that extend beyond the curriculum-related knowledge of particular disciplines. Standards for teachers' knowledge and skills may address teachers' knowledge of subject-matter content; teachers' knowledge of how to teach that content (content-specific pedagogy); teachers' knowledge of general pedagogical principles (e.g., language development, rewards and punishment); and teachers' ability to apply this knowledge and skill in a classroom setting (pedagogical performance). According to Bridges (1986), in practice, schools evaluate

already-employed teachers on five broad criteria: 1) knowledge of subject matter, 2) ability to impart knowledge, 3) ability to maintain classroom discipline, 4) ability to maintain a suitable classroom climate, and 5) ability to establish rapport with parents and students. As a general rule, however, school personnel do not have a specific understanding of the meaning of these criteria, and thus teachers are often unsure of what is expected of them, or how to improve their practice. In the school setting, determining teachers' competence is usually carried out through principals' or other supervisory staff observations, although that method is not the best way to obtain information on all five of the areas identified by Bridges. It is rare that passing a standardized test is a criterion for continued employment, but this has been done in several jurisdictions in the United States.

Teacher evaluations for purposes other than annual evaluation of already-employed teachers present different evaluation options. One example of a different approach has been created for use in certifying teachers as meeting very high professional standards and being outstanding practitioners. The NBPTS has designed complex and comprehensive assessments for experienced teachers that are tied to specific age levels of students, and to specific subject-matter areas. The NBPTS assessments are typically a mixture of assessment methods (heavily oriented toward performance assessment) that are specific to teaching in a particular context, such as early adolescent language skills. The assessments are external in the sense that they are judged by specially trained evaluators who are experienced educators and not part of the teacher's employment setting. NBPTS assessments are now available in certain age and core subject-matter areas that involve large numbers of teachers; others are still under development. In addition to assessment development, the NBPTS has also done a vast amount of research in a number of areas related to the use of simulations for the assessment of teachers (e.g., videotaped performance; the use of portfolios; assessment center exercises; and the impact of NBPTS assessments on teachers who participate in them).

Another model, originally developed for the assessment of beginning teachers, but now broadened to include experienced teachers as well, is that of The Praxis Series developed by Educational Testing Service (Dwyer, 1994; Dwyer and Ramsey, 1995; Dwyer and Villegas, 1993). The Praxis Series offers assessment and professional development activities that use a variety of assessment methods to collect data for personal and institutional decision making. Areas of assessment are:

- Basic enabling skills (reading, writing, and mathematics) required of all prospective teachers for success in teacher education and successful later practice.
- Subject-matter knowledge and content-specific pedagogical knowledge. These assessments use written, computer-based, oral, and other methods of data collection.
- Knowledge of basic pedagogical principles (using data-collection methods as described above).
- Application of knowledge and skills in the classroom. These assessments are predominantly performance assessments, and rely upon observation in the teacher's own classroom, teacher interviews, and data-gathering documents.

Level of accomplishment standards

In addition to specifying the "what" in teaching and learning standards, the question of "how much" must also be addressed. Curricular and teacher knowledge and skill standards often address this issue by implication, and less often address it directly. For example, when calculus is included as an element of mathematics standards, it implies a higher level of performance outcomes than standards that specify only that mathematical content is usually taught before calculus.

The issue of level of accomplishment standards is appropriately addressed independently of such considerations, however. Standards can be set at various levels (singly or in combination). For example, an overall standard may be set to determine, on a yes/no basis, a teacher's eligibility to teach, or to have an employment contract renewed. At a more detailed level, standards may be set to determine whether a teacher has adequate knowledge of a particular aspect of literature, or the ability to maintain discipline in the classroom.

Dwyer (1997) provides an overview of the basic principles used to establish such standards. The nature of cut scores[2] is not a matter of finding the "correct" score, but of the judgment of one or more empowered or authorized individuals about the question: "How much is enough?" The answer is not in the test, or in the method used to set the cut score. This is a question that can only be answered with respect to factors extraneous to the test itself, with reference to the context in which the test is used, and the judges' perceptions of that context.

In teacher evaluations, such standards can be highly controversial. The same standard can be seen simultaneously by different individuals as being either 1) so high that it represents an unconscionable barrier to the employment of qualified individuals who would make acceptable teachers, or 2) so low that it does a disservice to students by permitting unqualified individuals to teach them. Dwyer notes that this problem does not result from having picked the "wrong" cut score, or from technical deficiencies of the test on which the cut score is used. Instead, it results from clear disagreements about the relative importance of different kinds of classification error, and about the relative importance of educational policies such as teacher supply, raising educational standards, and concerns about fairness to teachers and students.

Three central, interrelated points about such standards are relevant to teacher evaluation:

1. All methods of setting cut scores depend on judgment. Judgments may be about people or about test questions (or other aspects of the test itself), but judgment with referents independent of the test are an intrinsic feature of any cut-score.
2. Setting cut scores will invariably lead to errors in classifying individuals as having met or not met the standard. Cut scores almost always impose external differentiations on a continuous distribution. Very few assess-

ments of any type can distinguish reliably between people with adjacent scores, yet applying a cut score, in effect, forces such a distinction.

3. No "true" cut score exists that can be found with the application of the correct method or a large enough sample of judges. Determining a cut score is not analogous to estimating a population parameter. Using larger samples of judges, or better trained judges, in studies to set cut scores will improve certain specific technical aspects (e.g., reducing sampling error) (Lawshe, 1975), but will not result in a cut score that is superior to those that might have been set with another selection of qualified judges, or with another equally justifiable methodology for selecting the cut score.

Opportunity-to-learn standards

In a context primarily concerned with standards for students, Banks (1997) notes the importance of opportunity in specifying standards. She notes that in setting the *Goals 2000* national educational standards in the United States, the types of standards considered included content standards, performance standards, assessment standards, and opportunity-to-learn standards. Opportunity-to-learn standards were the most controversial of the four. These standards address conditions in schools and communities that limit students' and teachers' ability to attain the other types of standards. Examples of such input variables include quality of school facilities, availability of teaching materials, and the level of teaching expertise available within the schools. Figure 1 shows the role of opportunity standards in the process of educational improvement. Despite the logic of having such standards, the criticisms that they imply of current mechanisms for supporting schools frequently make them unpalatable for political reasons. Critics of the use of such standards argue that they diminish the emphasis on achievement outcomes and accountability by inappropriately focusing on resources and input.
It should also be noted that in the special case of *licensing* beginning teachers, the concept of opportunity to learn is not relevant. In licensing contexts, the decision to be made is whether the teacher possesses the necessary knowledge and skills for eligibility for particular teaching situations. Because these situations are ones in which a major focus is the protection of the public from harm potentially done by an incompetent prospective teacher, that prospective teacher is being asked to demonstrate the possession of the stipulated knowledge and skills. The method by which one obtained these skills, or the circumstances of obtaining them, is not strictly relevant in this limited context. For further background on this issue, see Shimberg (1985).

PURPOSES FOR EVALUATING TEACHERS

The integrated process that I have just described of moving from goals to assessments can serve a number of purposes for teacher evaluation. Unfortunately, it is generally the case that the purpose of assessment is not universally clear to all interested parties when the decision is first made to create teacher assessments. In practice, there are often differing views about the purpose of assessments, even after they have begun to be used. Reaching a clear understanding of the reasons for assessing and the uses to be made of assessment data is a critical step in assessment design, but one that is often, mistakenly,

treated as self-evident. Thus, it is important to be explicit about the purposes teacher assessments are to serve.

Evaluation purposes

Different views of the purpose of educational improvement goals often imply different views of teacher evaluations. Such views of teacher evaluations may be quite different with respect to the substance of the assessments; the proponents and critics of the assessment; the fundamental rationales for the assessments; and the practical approaches to assessments (Dwyer and Stufflebeam, 1996). Some principal goals of teacher evaluations that are frequently observed in practice, and cited in the research literature, include:

- *Improvement of classroom teaching.* Professional educators are leading proponents of this view. It implies a continuum of educational development along which an individual may improve, a preference for formative rather than summative assessments, and a strong link to professional development activities. For further discussion of this purpose, see Hunter (1988); Duke and Stiggins (1990); and Shulman (1986).

- *Professional accountability and development.* Teachers and their professional associations are leading proponents of this view. It implies a strong view of teaching as a profession with its own standards, ethics, and intrinsic incentives for the committed individual. Although accountability is a key element of this view, the accountability is to the profession and its standards of practice and ethics, rather than to external entities such as employers or the state. For further discussion of this purpose, see National Board for Professional Teaching Standards (1991).

- *Administrative control.* School administrators are the leading proponents of this view. It implies regarding teaching as an employment situation that requires supervision and control of the teacher by the administrative unit. In the realm of public schools in the United States, the basis for this view is grounded in the protection of the public from negligent practice. For further discussion of this purpose, see Andrews (1985); and Redfern (1963, 1980).

- *Merit pay.* In this view, which can be seen as a subset of accountability or administrative control as described above, teachers are seen as needing the recognition and motivation that are provided by salary increases. Leading proponents of this view are the general public and the government officials who represent them. Proponents of this purpose often wish to use student achievement as the indicator of merit for which the increases in salary act as a reward. Implementation of this view may involve creation of a career ladder with a series of steps linked to performance. For further discussion of this purpose, see Webster, Mendro, and Almaguer (1993).

Accountability and educational improvement

Monitoring/accountability, instructional improvement, and program evaluation are the three most frequently cited purposes for teacher assessment in the United States. Although they are closely linked conceptually, assessments for these purposes are quite different in their design and in how the data from the assessments is used. For example, assessment for accountability and program evaluation may utilize sampling of individuals and assessment questions or exercises, rather than asking every individual to complete every part of the assessment. In contrast, assessments designed directly for instructional improvement are more appropriately given to every teacher or student, and should be very closely tied to the areas of curriculum that are most important for these students and their teachers.

Accountability is thus not a goal in itself, but rather a means to the end of educational improvement. Unless assessments of good and poor teacher performance are carefully planned, and the assessment system is fair in fact and in appearance, the assessment can actually subvert its intended aim of instructional improvement. The widespread use of assessments for teacher accountability in the United States is controversial. Research has demonstrated that without careful plans for the gathering and use of the data, results may be obtained that conflict with its original purposes. For example, teachers and administrators who do not value the content of assessments may resort to illicit means to obtain acceptable scores.

Both ethical and practical problems arise when the content of assessments is not closely linked to important aspects of teaching. In such circumstances, the time spent preparing for the assessments will, quite rightly, be considered wasted. When links between assessments and the broader educational goals and standards are absent, there is also reason for concern that "You get what you assess"—that is, that time and attention is devoted to those topics that appear on the assessments, at the expense of other important topics that do not. The format of assessments is sometimes cited as an issue in this respect. Excessive reliance on a single form of testing may result in teachers (or students) becoming adept at the kind of thinking required by that form of testing, at the expense of other forms of thinking.

Forging strong links between teacher accountability and educational improvement requires attention to the mechanisms and resources available for addressing problems when they are revealed by assessment data. The match between educational goals and standards on the one hand, and the format and content of teacher assessments on the other, helps educational policy makers and the public understand where problems exist, which may in turn suggest solutions. Different purposes for teacher assessment imply different standards, assessment methods, and emphases. An important factor in teacher evaluation for purposes related to administrative decisions such as accountability and merit pay, is the existence of actual or potential teacher shortages. Evaluations that exacerbate teacher shortages in critical areas will seldom be found useful. In the United States, such shortages have historically occurred in technical subjects such as mathematics, in which education competes, usually unsuccessfully, with industry for qualified

individuals. For discussions of these and related points on teacher supply and demand, see Murnane and Schwinden (1989) and Sedlak and Schlossman (1986).

Special considerations for beginning teacher assessments

Assessments of beginning and experienced teachers typically differ in the goals that the assessments are intended to serve, and the form that the assessments take. This is not to say that the two are unrelated; a hallmark of a carefully planned assessment of either type is the care with which it has been aligned with earlier or later stages in the teacher's career. A common vision of all teachers should govern the assessments and their contents at all stages.

Assessments of beginning teachers usually have the primary purpose of protecting the interests of the students and the public from the harmful effects of substandard teaching. Such assessments are typically geared toward ensuring that prospective teachers possess the knowledge and skills that qualify them for employment, but do not ordinarily guarantee teaching positions (Madaus and Mehrens, 1990; Mehrens, 1987; Rebell, 1990). The knowledge and skills covered are considered necessary, but not necessarily sufficient, indications of suitability for a particular teaching post.

In the United States, teacher licensing is typically under the authority of a state government, or an agency that it has empowered to perform the assessment function. Praxis assessments for licensing beginning teachers are an example of tests provided for this purpose. In contrast to assessments of experienced teachers, the assessment of beginning teachers covers enabling skills of reading, writing, and mathematics. Knowledge of subject matter and content-specific pedagogy in assessments of beginning teachers is congruent with expectations of experienced teachers. Proficiency in application of knowledge and skills in the classroom is similar in nature to that expected of experienced teachers, but at a lower proficiency level. One critical aspect of the assessment of beginning teachers is that assessments are used to make a dichotomous decision. Although they may serve other purposes as well, assessments of beginning teachers typically result in a yes/no decision about the prospective teacher's ability to move forward in the teaching profession. One measurement implication of this feature of beginning teacher assessments is that the assessment must be designed to facilitate this dichotomous decision, rather than to describe gradations of performance equally well along the entire continuum of teaching knowledge and skills. In practice, this means that the designers of the assessment must understand the qualities that differentiate an acceptable teacher candidate from an unacceptable candidate, and provide assessment tasks that will accurately allow for differentiations to be made in performance at this particular level, rather than across a wider range of performance.

Criticism of teacher evaluation

Teacher evaluation has been a particularly controversial area of educational testing almost since its inception in the early 1930s. Criticisms of teacher evaluations have differed during the years, and with differing purposes and methods of evaluation.

In general, criticisms of the evaluation of beginning teachers for entry into the profession have often stressed the fairness of instruments and procedures, especially with respect to the performance of prospective teachers who are minority group members (e.g., Darling-Hammond, 1986; Garcia, 1985; Haney, Madaus and Kreitzer, 1987; Hood and Parker, 1991). These criticisms charge that teacher evaluation practices are often carried out in inconsistent ways that violate the rights of those being assessed. Criticisms based on harm done to students by lack of teacher expertise are quite rare in this area.

Criticisms of the evaluation of practicing teachers for purposes of accountability, merit pay, etc., have also raised concerns about the actual or potential impact of teacher evaluation practices on prospective teachers' rights to teaching licenses or access to employment; on teacher educational institutions' curricula; and on the practice of teaching itself. The majority of the criticism appearing in the professional literature concerns the use of multiple-choice tests of subject-matter knowledge. Since the late 1980s, with increased use of performance assessments for both teachers and students, the criticism in the professional literature has increasingly focused on the practical aspects of performance assessments. Dwyer and Stufflebeam (1996) have identified eight major issues in teacher evaluations that relate primarily to the use of evaluations for teacher selection and employment:

- The use of unvalidated evaluation systems
- Insufficient use of professional standards for planning and improving educational systems
- Ineffectual choices of clear, valid, applicable criteria for assessing teacher performance
- The lack of techniques and materials to carry out the basic steps of teacher evaluation
- Lack of evaluation training for assessors
- Lack of guidance on evaluating teachers of special populations, e.g., students with handicaps, or students needing bilingual education
- Failure to consider the classroom and school context in carrying out evaluations
- Lack of theoretical grounding for evaluations

Despite widespread professional support for the validation of teacher evaluations (e.g., Linn et al., 1989; Madaus, 1990), most teacher evaluations used by schools have not been rigorously validated (Burry, Chisholm, and Shaw, 1990; Scriven, 1987; Streifer and Iwanicki, 1987).

Related to the last two issues in this list, some critics, who recognize the complexity of teaching, are specifically concerned about the feasibility of assessing pedagogical knowledge together with subject-matter understanding.

Multiple-choice questions that attempt to assess the application of pedagogical knowledge flounder on the question of context: what constitutes good teaching varies with the subject matter being taught, and with the background and individual characteristics of the students being taught. In the absence of very detailed information about the classroom context, which is infeasible in practical terms in the multiple-choice format, it is

impossible to conclude with any certainty what constitutes appropriate instruction. Assessment of teaching skills, as opposed to necessary subject-matter knowledge, thus needs to make use of more complex forms of data gathering.

Beginning teacher performance assessments created by states, although new relative to multiple-choice testing, have not escaped criticism. Such systems have been based on low-inference observations of teacher behavior; that is, the observation task is limited to clearly observable behaviors that require a low level of inference on the part of the raters, and thus can be expected to show a high level of inter-rater agreement. Although this high level of agreement is desirable, it is not so desirable if it comes at the cost of the validity of the ratings. By their nature, low-inference ratings cannot assess important but unobservable facets of teaching, such as teacher decision making. Many critics of performance-based teacher evaluations have pointed to the failure of low-inference systems—both those based on simulations and those based on direct classroom interactions—to take the classroom context into account as part of the assessment.

Other criticisms of teacher evaluation concern the impact of teacher evaluation on teacher education curricula (Milner, 1991), and on the profession of teaching itself (Darling-Hammond, 1986). The preponderance of this criticism concerns the use of standardized multiple-choice tests of subject matter and pedagogical knowledge, rather than performance assessments of the ability to apply this knowledge in the classroom. An important issue in understanding criticism of teacher evaluations and in formulating policies that avoid legitimate criticisms, is making a consistent distinction between the two types of testing. Unfortunately, assessments of basic enabling skills and knowledge of subject matter, content-specific pedagogy, and general pedagogical skills are too often treated as interchangeable with assessments of classroom performance. Inferences made on the basis of a teacher's knowledge pertain to a necessary *but insufficient* condition for effective classroom teaching. Failure to make this simple but critical distinction muddies many debates about teacher evaluation policies and practices. Many of the documented shortcomings of teacher evaluations can be traced to lack of resources such as appropriate training procedures for evaluators, evaluation instruments, and information from validation studies. Other criticisms, however, are the result of the current state of teacher evaluation theory and practice. The absence of theoretical grounding is a significant problem for many systems in evaluating both beginning and experienced teachers (Scriven, 1988a, 1988b; Wise, Darling-Hammond, McLaughlin, and Bernstein, 1984). Unfortunately, the teaching research literature offers little guidance on how to use their theory to develop and implement better teacher evaluations (Dwyer, 1994).

DEFINING TEACHING

A major challenge in creating a well-articulated set of teacher evaluations is the task of determining what constitutes good teaching, and what standards should be used to determine if it has been measured effectively. The adequacy of measuring what should be taught cannot be meaningfully determined without articulating explicit standards, in a comprehensive frame of reference that encompasses traditional issues of concern to

education, psychology, and measurement. This evaluation must also be set in a context relative to a particular view of teaching and learning. All these factors form an appropriate part of a concept of teaching that is used to guide decisions about standards and assessments. Conclusions about the quality of assessments, in turn, must be referenced to the concept of teaching and the purposes of the assessments. Recent developments in measurement strongly support this point of view.

Modern validity theory, in its emphasis on a broad context for establishing the validity of assessments, provides conceptual guidance in considering the validity of assessments in complex activities such as teaching, and on the technical aspects of determining their content. Such work on validation theory highlights the measurement implications of the interconnections that exist within the whole education system of which assessment is a part, and the value inherent in direct measurement of performance where this is feasible. With the emphasis on context and consequences, assessments are examined, in broad terms, to determine their benefit to the educational system. For example, in teacher assessment, the broad context of the assessment of a beginning teacher would include such factors as school district, state, or federal policies that affect the teacher's practice; the school and classroom setting in which the teaching occurs; material resources available to the teacher; the nature of the curriculum and lessons to be taught; the characteristics and prior knowledge of the students; and the quality of the teacher's preparation to teach. The consequences of the use of teacher assessments would similarly focus on the impact of the use and outcomes of the assessment on educational policy (Are new policies created, based on the assessment outcomes? If so, can these new policies be shown to be beneficial to students?); materials and facilities available to teachers and students (Are material needs identified through the assessment, then filled? Does the assessment process divert attention from material needs, so that such needs worsen?); the quality of teacher preparation (Does the assessment provide information to teacher educational programs that helps them improve their program?); the quality of the curriculum (Do the assessment process and results lead to identification of areas of the curriculum that work well, and other areas that do not?); and the learning of the students (Does teacher assessment lead to improved learning, or does it interfere with classroom practice so that students learn less?).

Although the theoretical basis for a broad, construct view of validity, including its emphasis on consequences of assessment, is now widely accepted, and is in fact codified in the most recent revision of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1985), there is still a considerable gap between the literature on validity and the diverse and challenging set of issues faced in developing teacher assessments, particularly when they take place in high-stakes environments, and when multiple means of gathering data are used. For example, researchers and developers in this area lack definitive guidance from the literature on such important issues as classroom subject matter and human variables in teacher assessments (how to deal with contextual differences); and on the validity implications of creating assessments that take as their starting point the view that learning is an active process of constructing meaning from prior experiences (assessments with a constructivist founda-

tion). These issues are critical to determining the criteria by which teaching performance will be assessed, but a leap is required to bridge the solid theoretical base in the research literature to actual research and development practice.

The body of literature on validity, standards, expectations, and aspirations—while highlighting important conceptual and practical issues—creates a heavy burden of interpretation and extrapolation for developers of teacher assessments, particularly those that include attention to actual classroom performance. There is clearly general agreement in this literature that assessment developers and evaluators should take the broad consequences of assessment into account; should incorporate elements of context into the assessment process; should focus explicitly and accurately on the knowledge and skills about which one wishes to draw inferences; and should include in the assessments the full range of content about which inferences are to be drawn. Despite this high level of agreement in principle, a great deal of discretion and responsibility is necessarily left to individual creators of assessments and to those who evaluate their efforts. The basis for their decisions, if carefully considered, can be a major asset in aligning teacher evaluations with other elements of the educational system in order to improve educational outcomes and meet other important goals.

Viewed as a process, the major elements in the process of defining teaching for purposes of assessment include:

- Determining a guiding concept of teaching
- Developing a comprehensive methodological plan for defining teaching
- Linking the definition of teaching to assessment practices

Guiding conception of teaching

It is critically important to the development of effective teacher evaluations to formulate a concept to guide teaching and learning that explicitly recognizes the *connection* between teaching and learning. It is impossible to discuss what is fundamental about one without considering the other. An effective concept to guide teaching should be:

- Explicit and clear about where criteria come from, whose values they are in agreement with, and what value positions they reject
- Clear about issues that are inherently matters of teaching style with those that are important matters of substance (Scriven, 1988a, 1988b)
- Should lead to inferences about both the content of the assessments and the methods used to collect data.

For these reasons, the concept to guide teaching should be articulated early in the development process, before final design decisions are made.

Again, the Praxis Series is an example of the development of such a guiding concept. This concept has been formalized by Dwyer and Villegas (1993). In the development of Praxis, emphasis was placed on teacher decision making and on the importance of the

student, school, and curricular context in evaluating that decision making. This point of view strongly implies the value of data-gathering in the actual classroom setting, as opposed to simulations. A second implication is that the assessments should include opportunities for the assessor and assessee to interact regarding the teaching event that is being considered. This will help the assessor to fully understand the decisions on which the observable teacher actions were based. A third implication is that because there is no "one right answer" to the question of what constitutes good teaching (because teaching is seen as inherently context sensitive), the scoring of the assessments must allow for multiple forms of acceptable "answers," while clearly articulating what constitutes unacceptable professional practice.

A fourth implication related to the concept of teaching as a complex cognitive activity is that the assessment process, relying as it does on consideration of a complex set of data, will require substantial professional judgment to implement. Assessors must thus be experienced professionals, who have been trained to reach a common understanding of the assessment criteria and other considerations for applying them.

This concept of teaching and learning, and the criteria that flow from it, has strong links to psychological, educational, and measurement theory and practice. It specifies a cognitively and behaviorally complex target performance, and provides a framework for examination of the impact of the assessments on the educational system of which it is a part (students; the teachers being assessed; the teaching profession; teacher education and staff development).

In the case of Praxis, the following values were also explicitly identified as part of developing the guiding concept, and were used to guide the actual assessment development:

- Commitment to the equitable treatment of teachers
- Standards and assessment techniques that deal with both teachers' actions and teachers' decision making
- Specification of certain teaching practices as unacceptable in any context, while allowing for many different modes of acceptable practice, allowing the creation of specific and meaningful standards of teaching knowledge and practice
- Creating a positive learning experience for both the teacher being evaluated and the assessor

Specifying a methodology for defining teaching

Figure 2 shows a sample methodology for defining the domain of teaching for the classroom performance for the beginning teacher.

This methodology was used in the creation of The Praxis Series assessments. Note that this is an iterative process that integrates data from theory through formal research studies and literature reviews; and practice through teacher survey, participation, and iterative field work.

The following section describes significant issues involved in implementing this methodology and in developing teacher assessments based on this analysis.

ISSUES IN DEVELOPING TEACHER EVALUATIONS

Selecting assessment methods

The best strategy for selecting methods for teacher assessment can be summed up very simply: identify what is important to know about the teacher, and then determine the optimal procedure for obtaining the necessary information. Possibilities include multiple-choice testing, other forms of written assessments, and many types of performance assessments. In most cases, adequate coverage of the necessary knowledge and skills will require a combination of assessment techniques. Criteria for selecting the best method for assessing particular content and skill areas should be comprehensive and include:

- Consistency with the guiding concept of teaching
- Logical relationship to the target area of knowledge or skill
- Positive educational impact
- Operational and economic feasibility
- Demonstrable fairness to test-takers

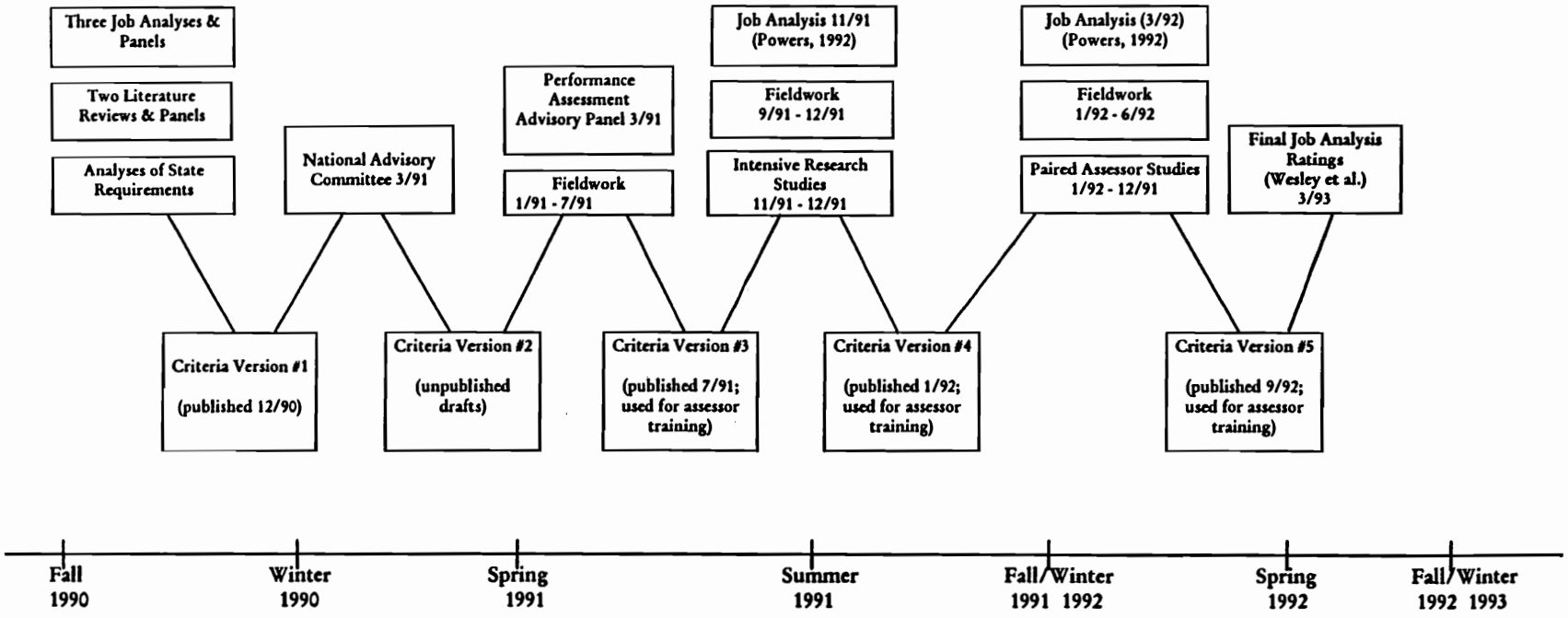
As noted above, valid and high-quality assessments should result in a positive educational impact as articulated in Messick's (1989) view of construct validity. In this view, a valid test leads to good educational practice and worthwhile learning, while an invalid test leads to suboptimal skills development.

Fairness to teachers is a key criterion for judging any prospective assessment methodology. Fairness can be considered on the basis of characteristics such as gender, race, or ethnicity; or on other factors such as educational background, relevant life experiences, and familiarity with the subject matter and students to be taught.

Other considerations regarding fairness to the individual are reliability of the assessments, which must be judged by measures appropriate to each method used in the assessment. Different forms of assessment require different methods for establishing reliability as well as differing standards for designating appropriate levels of reliability. Specifically, the methods and levels used to judge multiple-choice tests are not necessarily appropriate to apply to other forms of assessment.

An important consideration in choosing an assessment method is that the method must be operationally and economically feasible for large-scale use. The short-term economy and efficiency of multiple-choice testing are obvious. It is advisable, however, in evaluating new methods of assessment, to compare the total effect of the various forms of assessment. Within carefully defined limits, the added value of gains in validity, directness of measurement, candidate motivation, or other measures of assessment quality can offset increased costs and operational complexity in other forms of assessment. Central test quality standards such as reliability, validity, and fairness are still relevant to the various forms of assessment that are chosen.

Figure 2. Praxis III: Evaluation of Classroom Performance: Framework for Establishing Criteria



The economics of performance assessments, and constructed response testing in general, present formidable challenges. Compared to multiple-choice testing, these methods of assessment are inherently labor intensive, and thus more expensive to implement. For this reason, what is measured with performance assessments and constructed response methods must be more meaningful to justify the additional required expenditures. Constructed response testing, taken as a whole, often serves as an educational experience both for those doing the assessing and for those who are the objects of the assessment. This view of evaluation as an inservice activity is often part of the rationale for additional expenditures for performance assessments.

In the case of the Praxis assessments, a variety of assessment methods were selected. The Praxis I, Assessments of Basic Enabling Skills of Reading, Writing, and Mathematics are used for those people entering teacher education and are assessed through a combination of computer-based, multiple-choice, open-ended, and essay questions. Praxis II, Assessments of Subject Matter Knowledge, differ in methodology according to the discipline. Almost all the Praxis II tests contain some multiple-choice and open-ended questions, but certain aspects differ greatly. (An assessment of language teachers, for example, uses audiovisual response modes to assess receptive and productive aspects of communicative competency. An assessment of physical education teachers uses video stimuli to assess important aspects of body movements.) The Praxis III, Assessments of Classroom Performance, use a combination of interviews, observations, and written documentation to assess teacher effectiveness in an actual classroom setting.

Lead or lag?

Developers of the teacher evaluation systems are also faced with what is often called the "lead/lag" dilemma. In fairness to the teachers who are being evaluated, the criteria used to judge them must reflect currently acceptable professional practice. This also ensures that evaluations are logically consistent with the purposes of the assessments. A competing value, however, is that given the long lead-time for developing high-quality assessments, and the likelihood that they will continue to be used for a number of years, it is also important not to create assessments that will be, in effect, obsolete before they are completed, or that will encourage continuation of teaching practices that are even now only marginally acceptable to the profession.

Central to this issue is that acceptable professional practice is not a static concept; new knowledge about teaching is created on a daily basis. In evaluating whether a particular aspect of teaching can be considered to be supported by research, it is therefore necessary to make a number of complex judgments about the status of the research and to take into consideration the professional consensus about future trends in that area. It is also necessary to consider how definitive a research area is at the time. For example, the area of teacher behavior and its links to student learning has been extensively researched. In particular subareas, the domains are well mapped. Well-designed studies are numerous, and in some cases, definitive conclusions have been reached.

In contrast, the area of teacher cognition and its links to student learning is still relatively young and in a state of flux. Although the importance of this research domain to teaching practice is not generally in dispute, the nature of teacher cognition itself is still to some extent uncertain. A number of important principles, although logically well-articulated and convincingly demonstrated in the best research studies, have not yet been widely replicated. Development of teacher assessments must make judgments based on such observations about what to include for the purposes of teacher evaluations. A complicating factor is that research on teacher behavior and teacher cognition tend to utilize different methodologies, thus creating another difficulty in evaluating the newer research by traditional standards.

In the case of the development of the Praxis assessments, the majority of the content was drawn from current practice, but certain aspects were judged to be part of current trends. For example, in Praxis III classroom performance assessments, greater emphasis was placed on the importance of teachers working with each other and with students' families than surveys indicate is the current practice. The inclusion of this facet of teaching was based on professional consensus as well as research that such interactions are beneficial to student learning.

Practical and theoretical emphases

Another dilemma faced by the developers of any teacher evaluation system is the relative merit of the theoretical and practical knowledge of teaching. Sternberg and Wagner (1993) have drawn a useful distinction between academic and practical problems that may be of interest with respect to teacher evaluation. According to Sternberg and Wagner, academic problems tend to a) be formulated by other people, b) be well-defined, c) be complete with regard to the information needed to solve them, d) possess only a single correct answer, e) possess only a single method of obtaining the correct answer, f) not be embedded in ordinary experience, and g) be of little or no intrinsic interest. In contrast, practical problems tend to a) require problem recognition and formulation, b) be ill-defined, c) require information-seeking, d) possess multiple acceptable solutions, e) allow multiple paths to solution, f) be embedded in and require prior everyday experience, and g) require motivation and personal involvement. It is clear that in Sternberg and Wagner's terms, comprehensive teacher evaluations address the practical problems that teachers must solve, rather than academic problems.

Moreover, as noted above, it is important for both practical and theoretical reasons that the criteria on which teacher assessments are based, and their organizing framework, correlate with teachers' own understandings of their work. At the same time, the criteria should also be based on educational and psychological theory to establish a logical framework. There should also be increased generalization across teaching contexts, and an increased probability of standing the test of time in actual classroom use. Again, resolving this dilemma is not a mechanistic process. Seeking the input of practicing teachers and educational theoreticians to review and revise the evaluation materials until they are broadly perceived as acceptable from both points of view has been a successful strategy in the development of the Praxis assessments.

From whose perspective should the knowledge base be considered?

As noted above, principal sources of data for the knowledge base about teaching include practicing teachers themselves and the research/theoretical perspective. Additional sources may consist of regulations that replace governing schools and teachers. Each of these perspectives does not simply provide a different view of the same phenomena; each asks different questions, employs different methods to reach conclusions, and has a set of different, although often related, concerns about the meaning and use of knowledge about teaching. These views represent fundamentally different paradigms, in the sense that basic assumptions, methodologies, and values differ.

It is therefore not an algorithmic or mechanical process to arrive at criteria that incorporate data from these three sources. As an example of one solution to this problem, the developers of The Praxis Series carried out an iterative procedure of creating draft criteria, then presenting them for review to representatives of these three main points of view. Reviewers and panelists were asked, in essence, if the draft criteria represented the knowledge base for teaching as they understood it. The criteria underwent a number of major revisions as a result of this process. With each of these major revisions, increasingly large cycles of fieldwork were undertaken to provide additional data. The Praxis III criteria were considered final when each of the constituencies consulted were satisfied that their major concerns had been addressed.

Scope of the assessment criteria

Two interrelated issues that might be characterized as issues related to the scope of the criteria need to be addressed in the process of translating knowledge of the domain of teaching into specific teacher evaluation criteria: finding the appropriate "size" or level of generality for the criteria, and determining the range of teaching contexts to which the criteria apply.

The enormous variability among classroom contexts (Shulman, 1988, 1988a; Stodolsky, 1988) poses significant challenges for any teacher assessment effort. In the case of the Praxis III, Classroom Performance Assessments, the teaching criteria were designed as *aspects* of teaching; that is, as principles to be applied in a wide range of teaching contexts (including variability in subject matter and grade level taught, teaching style, and students' individual and background characteristics) rather than as specific "rules" to be followed or behaviors to be demonstrated. This implies a need for attention to consistency among the criteria.

A major issue in being able to design practical assessments that can accommodate a wide range of classroom contexts has been determining the right size concept for the criteria. Some researchers (Kagan, 1990; Katz and Raths, 1985) have called this challenge the "Goldilocks Principle." In teacher evaluation, this principle means that if the criteria are too big, i.e., too vague and general, then meaningful standards are difficult to develop and to apply fairly; assessors cannot apply a consistent set of judgments to the assessment process. On the other hand, if criteria are too small, i.e., too specific,

they can be judged with great consistency, but they will not capture the essence of good teaching and may promote a fragmented, cookbook approach to teaching. Thus, the assessment criteria, like the bears' porridge that Goldilocks tasted, must be "just right" if they are to meet the goals of fair assessment and improving teaching practice. As a practical matter, finding the level of specificity that is "just right" involves many iterations of fieldwork and analysis, such as those shown in Figure 2.

In the reports of the fieldwork for the Praxis assessments, there are numerous instances of these experiences leading the developers to conclude, for example, that what had been a single criterion ought to be divided into two separate criteria to help assessors better understand how a particular aspect of teaching is actually played out in the classroom and help them recognize evidence related to this aspect of teaching when they see it.

Such fieldwork also enables the assessors and developers to see how the criteria relate to each other in practice, and to use this information as the basis for making changes in how the criteria are organized, ordered, and described. Organizing and wording the criteria so that they are clear and logical, from the point of view of those who use them, should be given a high priority in the development work.

Assessors and judgment

As noted above, judgment plays an important part in developing any instrument for the evaluation of teachers. In multiple-choice testing, judgment is exercised at the level of deciding such issues as what to test, how questions should be presented, what constitutes a correct answer, and the number of questions that need to be answered correctly in order for a candidate to be considered successful. Thus, even in this familiar form of "objective" testing, judgment plays an important role, although not in the scoring itself. Performance assessments, on the other hand, incorporate judgment in many ways that are similar, but also in some ways that are different. When using performance assessments to evaluate teachers, it is critical to understand the role played by the professional judgment of those who assess them.

The specification of assessment criteria is an important part of developing any performance assessment, but the success of the effort as a whole can only be evaluated in light of the ability of assessors to use these criteria to reach technically and professionally defensible conclusions.

Unlike traditional multiple-choice testing, where the great preponderance of the professional judgment comes into play during the preadministration phase of test development, professional judgment in performance assessments is required in both the development and the use phases of the assessment. The quality of this professional judgment affects many important aspects of the assessment's validity, including, but not limited to, fairness, cognitive complexity, and construct representation. It is also related to concerns for generalization, although the concerns are not the same for natural classroom performance assessment as they are for simulation-based assessment.

In classroom performance assessment, the generalization concern is to other teaching events, not to other aspects of teaching. In this sense, the teaching events are analogous to exercises or tasks in other types of performance assessments. An example of this can be found in the Praxis III, Classroom Performance Assessments (see Table 1 for a description of the criteria for assessment in The Praxis Series).

In the Praxis model, the "scoring" of the "tasks" is held constant via the criteria and their associated scoring rules. In assessing live teaching performance, variability across "tasks" is a natural and acceptable phenomenon, and thus inferences based on a given set of teaching events are expected to be general to an intrinsically variable universe of teaching events that defines the construct. Generalization across tasks is therefore not problematic in the same sense as when tasks are seen as partial or indirect instantiations of the construct.

As noted above, the Praxis III, Classroom Performance Assessments criteria are intended to be construed as interrelated aspects of a complex performance, not as functionally independent entities. As such, one would not aim to generalize from one aspect of teaching to another as evidence of validity, but rather to investigate the patterns of ratings given across occasions, and within a single occasion by two or more assessors (assuming that occasions are expected to be highly variable, relative to within- or across-assessor variability).

The assessment criteria do not stand alone because they are aspects of teaching, and not particular behaviors. They must be interpreted in light of the actual classroom context, which includes both the students and the subject matter being taught. The criteria serve as the guide for structuring assessors' judgments, ensuring that a common frame of reference rather than personal preference is the basis of the assessors' conclusions and ratings. Assessors' professional judgments are thus the cornerstone of the defensibility of the ratings of the beginning teacher.

Using the methods described above, assessors gather and organize data that affect each criterion; make critical judgments about the importance of the evidence and its relevance to a particular criterion; then reach a conclusion about the beginning teacher's performance level on each criterion based on this evidence. Assessors document these judgments by citing specific evidence and linking it to a rating scale that describes increasingly proficient levels of performance with respect to each of the criteria. Legitimacy of the assessment process is thus based on the quality of this argument (structured, documented, professional judgment) rather than on a purported absence of human decision making (objectivity). Through special studies (such as paired-assessor comparisons), field work in a variety of teaching settings, and operational use, methods of data gathering other than those we now use may be found to result in better measurement—that is, in more accurate or detailed judgments of the criteria, in better documentation, or in more positive effects on the system of which the assessment is a part. In determining the value and validity of the assessments, however, the data-gathering methods themselves are clearly subordinate to the quality of the criteria and to the assessors' judgments and systematic application of the assessment procedures.

CONCLUSION

Teacher evaluation is a key element in any effort to improve education. Well-conceived and executed systems for teacher evaluation are an integral part of the educational system, and support its main goals. Poorly conceived and executed systems may have a deleterious effect on the very system they are intended to improve. Thus, careful attention to the design of teacher assessments, monitoring for adherence to the plans during the implementation phase, and systematic evaluation of the entire assessment system and its consequences throughout its operation are essential factors in deriving the expected benefits from teacher assessments.

Educators have historically had to struggle with the tension among the competing concerns of equity, excellence, and efficiency. At various times, the balance has shifted to concern about one of these at the expense of the others. In order to best serve the needs of education and all those it affects, however, we need to strive to see beyond our own daily and individual concerns, and to maintain a focus on the larger picture in order to achieve the best possible balance of these three most urgent concerns.

Table 1: Assessment Criteria for the Praxis Series: Classroom Assessments for Beginning Teachers

Domain A: Organizing Content Knowledge for Student Learning. Knowledge of the content to be taught underlies all aspects of good instruction. Domain A focuses on how teachers use their understanding of students and subject matter to decide on learning goals; to design or select appropriate activities and instructional materials; to sequence instruction in ways that will help students to meet short and long-term curricular goals; and to design or select informative evaluation strategies. All of these processes, beginning with the learning goals, must be aligned with each other, and because of the diverse needs represented in any class, each of the processes mentioned must be carried out in ways that account for the variety of knowledge and experiences that students bring to class. Therefore, knowledge of relevant information about the students themselves is an integral part of this domain.

Domain A is concerned with how the teacher thinks about the content to be taught. This thinking is evident in how the teacher organizes instruction for the benefit of the students. The primary sources of evidence for this domain are the Class Profile, Instruction Profile, and Preobservation Interview. The classroom observation may also contribute to assessing performance in this area.

Assessment Criteria for Domain A:

A1: Becoming familiar with relevant aspects of students' background knowledge and experiences

A2: Articulating clear learning goals for the lessons that are appropriate for the students

A3: Demonstrating an understanding of the connections between previously-learned, current, and remaining content

A4: Creating or selecting teaching methods, learning activities, instructional materials, or other resources that are appropriate for the students, and that are aligned with the goals of the lesson

A5: Creating or selecting evaluation strategies that are appropriate for the students and that are aligned with the goals of the lesson

Domain B: Creating an Environment for Student Learning. Domain B relates to the social and emotional components of learning as prerequisites to, and context for, academic achievement. Thus, most of the criteria in this domain focus on the human interactions in the classroom, on the connections between teachers and students, and among students. Domain B addresses issues of fairness and rapport, of helping students believe that they can learn and can meet the challenges of establishing and maintaining constructive standards for behavior in the classroom. It also includes the learning "environment" in the most literal sense—the physical setting in which teaching and learning take place.

A learning environment that provides both emotional and physical safety for students is one in which a broad range of teaching and learning experiences can occur. Teachers must be able to use their knowledge of their students in order to interpret their students' behavior accurately and respond in ways that are appropriate and supportive. When they do so, their interactions with students consistently foster the students' sense of self-esteem. In addition, teachers' efforts to establish a sense of the classroom as a community with clear standards should never be arbitrary; all behavioral standards and teacher-student interactions should be grounded in a sense of respect for all members of the classroom community.

Evidence for the criteria in Domain B is drawn primarily from the classroom observation; supporting evidence may be drawn from both the pre- and postobservation interviews. The Class Profile provides contextual information relevant to the criteria comprising this domain.

Assessment Criteria for Domain B:
B1: Creating a climate that promotes fairness
B2: Establishing and maintaining rapport with students
B3: Communicating challenging learning expectations to each student
B4: Establishing and maintaining consistent standards of classroom behavior
B5: Making the physical environment as safe and conducive to learning as possible

Domain C: Teaching for Student Learning. This domain focuses on the act of teaching and its overall goal: helping students to connect with content. As used here, "content" refers to the subject matter of a discipline and may include knowledge, skills, perceptions and values in any domain: cognitive, social, artistic, physical, and so on. Teachers direct students in the process of establishing individual connections with the content, thereby devising a good "fit" for the content within the framework of the students' knowledge, interests, abilities, and cultural and personal backgrounds. At the same time, teachers should help students move beyond the limits of their current knowledge or understanding. Teachers monitor learning, making certain that students assimilate information accurately and that they understand and can apply what they have learned. Teachers must also be sure that students understand what is expected of them procedurally during the lesson and that class time is used to good purpose.

Most of the evidence for a teacher's performance with respect to these criteria will come from the classroom observation. It may be augmented or illuminated by evidence from the pre- and postobservation interviews, the Instruction Profile, and the Class Profile.

Assessment Criteria for Domain C:
C1: Making learning goals and instructional procedures clear to students
C2: Making content comprehensible to students
C3: Encouraging students to extend their thinking
C4: Monitoring students' understanding of content through a variety of means, providing feedback to students to assist learning, and adjusting learning activities as the situation demands
C5: Using instructional time effectively

Domain D: Teacher Professionalism. Teachers must be able to evaluate their own instructional effectiveness in order to plan specific future lessons for particular classes and to improve their teaching over time. They should be able to discuss the degree to which different aspects of a lesson were successful in terms of instructional approaches, student responses, and learning outcomes. Teachers should be able to explain how they will proceed to work toward learning for all students. The professional responsibilities of all teachers, including beginning teachers, also include sharing appropriate information with other professionals and with families in ways that support the learning of diverse student populations.

Assessment Criteria for Domain D:
D1: Reflecting on the extent to which the learning goals were met
D2: Demonstrating a sense of efficacy
D3: Building professional relationships with colleagues to share teaching insights and to coordinate learning activities for students
D4: Communicating with parents or guardians about student learning

NOTES

[1] The term evaluation is used throughout this paper to indicate a range of activities undertaken to gather information for the purposes of decision making. The term is used without reference to distinctions among such purposes, and includes, but is not restricted to, discussions of specific assessment techniques. The term assessment will be used to refer to these specific data-gathering techniques. In the United States at the present time, the use of these two terms is somewhat controversial, with evaluation referring primarily to activities carried out with practicing teachers by those who employ them, for the purposes of making individual personnel decisions.

[2] The term *cut score* will be used in this section to refer to any standard indicating a level of accomplishment. The use of this term is applicable to any type of assessment in which these evaluated are to be divided into groups for decision making.

REFERENCES

American Educational Association, American Psychological Association, and National Council on Measurement in Education. (1985). *Standards for Educational and Psychological Testing.* Washington, DC: American Psychological Association.

Andrews, H.A. (1985). *Evaluating for Excellence: Addressing the Need for Responsible and Effective Faculty Evaluation.* Stillwater, OK: New Forums Press.

Banks, C.A.M. (1997). "The Challenges of National Standards in a Multicultural Society." *Education Horizons,* 75, 3: 126-132.

Berk, R.A. (1988). Fifty Reasons Why Student Achievement Gain Does Not Mean Teacher Eeffectiveness. *Journal of Personnel Evaluation in Education,* 1, 345-363.

Bridges, E. (1986). *The Incompetent Teacher: The Challenge and the Response.* Philadelphia: Falmer Press.

Burry, J.A., Chissom, B.S., and Shaw, D.G. (1990, March). *Validity and Reliability of Classroom Observations: A Paradox.* Paper presented at a meeting of the National Council on Measurement in Education, Boston.

Claxton, C., Murrell, P.H., and Porter, M. (1987). "Outcomes Assessment." *AGB Reports,* 29, 5: 32-35.

Cole, N.S., and Moss, P.A. (1989). "Bias in Test Use." In R.L. Linn (ed). *Educational Measurement* (3rd ed., pp. 201-219). New York: Macmillan.

Darling-Hammond, L. (1986). "A Proposal for Evaluating in the Teaching Profession." *Elementary School Journal,* 86: 531-569.

Duke, D., and Stiggins, R.J. (1990). "Beyond Minimum Competence: Evaluation for Professional Development." In J. Millman and L. Darling-Hammond (eds). *The New Handbook of Teacher Evaluation: Assessing Elementary and Secondary School Teachers* (pp. 116-132). Newbury Park, CA: Sage.

Dwyer, C.A. (1993b). "Innovation and Reform: Examples from Teacher Assessment." In R.E. Bennett and W.C. Ward (eds). *Construction Versus Choice in Cognitive Measurement* (pp. 265-289). Hillsdale, NJ: Lawrence Erlbaum.

Dwyer, C.A. (1993c). "Teaching and Diversity: Meeting the Challenges for Innovative Teacher Aassessments." *Journal of Teacher Education,* 44(2): 119-129.

Dwyer, C.A. (1994). "Criteria for Performance-Based Teacher Assessments: Validity, Standards, and Issues." *Journal of Personnel Evaluation in Education,* 8: 135-150.

Dwyer, C.A. (1996). "Cut-Scores and Testing: Statistics, Judgment, Truth, and Error." *Psychological Assessment,* 8 (4).

Dwyer, C.A., and Ramsey, P.A. (1995). "Equity Issues in Teacher Assessment." In M.T. Nettles and A.L. Nettles (eds). *Equity and Excellence in Educational Testing and Assessment.* Boston: Kluwer Academic.

Dwyer, C.A. and Stufflebeam, D. (1996). "Teacher Evaluation." In D. Berliner and R. Calfee (eds). *Handbook of Educational Psychology* (pp. 765-786). New York: Macmillan.

Dwyer, C.A. and Villegas, A.M. (1993). *Guiding Conceptions and Assessments Principles for The Praxis Series: Professional Assessments for Beginning Teachers* (RR 93-17). Princeton, NJ: Educational Testing Service.

Evertson, C., and Green, J. (1986). "Observation as Method and Inquiry." In M. Wittrock (ed). *Handbook of Research on Teaching* (3rd ed., pp. 162-213). New York: Macmillan.

Hunter, M. (1988). "Effecting a Reconciliation Between Supervision and Evaluation." *Journal of Personnel Evaluation in Education,* 1: 275-279.

Garcia, P.A. (1985). *A Study of Teacher Competency Testing and Tests Validity with Implications for Minorities: Final Report.* (NIE Grant No. NIE-G-85-0004). Edinburgh, TX: Pan American University.

Glass, G.V. (1990). "Using Student Test Scores to Evaluate Teachers." In J. Millman and L. Darling-Hammond (eds). *The New Handbook of Teacher Evaluation* (pp. 191-215). Newbury Park, CA: Sage.

Haney, W., Madaus, G., and Kreitzer, A. (1987). "Charms Talismanic: Testing Teachers for the Improvement of American Education." In E.Z. Rothkopf (ed.), *Review of Research in Education* (Vol. 14, pp. 169-238). Washington, DC: American Educational Research Association.

Hood, S., and Parker, L. (1991). "Minorities, Teacher Testing, and Recent U.S. Supreme Court Holdings: A Regressive Step." *Teachers College Record,* 92: 603-618.

Joint Committee on Standards for Educational Evaluation. (1988). *The Personnel Evaluation Standards.* Newbury Park, CA: Sage.

Kagan, D.M. (1990). "Ways of Evaluating Teacher Cognition: Inferences Concerning the Goldilocks Principle." *Review of Educational Research,* 60: 419-469.

Lawshe, C.H. (1975). "A Quantitative Approach to Content Validity." *Personnel Psychology,* 28: 563-575.

Linn, R.L., Baker, E.L., and Dunbar, S.B. (1991). "Complex, Performance-Based Assessment: Expectations and Validation Criteria." *Educational Researcher,* 20 (8): 15-21.

Madaus, G.F. (1990). "Legal and Professional Issues in Teacher Certification Testing: A Psychometric Snark Hunt." In J.V. Mitchell, S.L. Wise, and B.S. Plake (eds). *Assessment of Teaching: Purposes, Practices, and Implications for the Profession* (pp. 209-259). Hillsdale, NJ: Lawrence Erlbaum.

Madaus, G. and Mehrens, W.A. (1990). "Conventional Tests for Licensure." In J. Millman and L. Darling-Hammond (eds). *The New Handbook of Teacher Evaluation: Assessing Elementary and Secondary School Teachers* (pp. 257-277). Newbury Park, CA: Sage.

Madaus, G.F., and Pullin, D. (1987, September). "Teacher Certification Tests: Do They Really Measure What We Need to Know?" *Phi Delta Kappan,* 69(1): 31-38.

Mehrens, W.A. (1987). "Validity Issues in Teacher Licensure Tests." *Journal of Personnel Evaluation in Education,* 1: 195-229.

Messick, S. (1989). "Validity." In R.L. Linn (ed). *Educational Measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Miller, M.D., and Legg, S.M. (1993). "Alternative Assessment in a High-Stakes Environment." *Educational Measurement: Issues and Practice,* 12 (2): 9-15.

Millman, J. (1981a). *Handbook of Teacher Evaluation.* Beverly Hills, CA: Sage.

Millman, J. (1981b). "Student Achievement as a Measure of Teacher Competence." In J. Millman (ed). *Handbook of Teacher Evaluation* (pp. 146-166). Beverly Hills, CA: Sage.

Milner, J.O. (1991). "Suppositional Style and Teacher Evaluation." *Phi Delta Kappan,* 72: 464-467.

Moss, P.A. (1992). "Shifting Conceptions of Validity in Educational Measurement: Implications for Performance Assessment." *Review of Educational Research,* 62: 229-258.

Murnane, R.J., and Schwinden, M. (1989). "Race, Gender, and Opportunity: Supply and Demand for New Teachers in North Carolina, 1975-1985." *Educational Evaluation and Policy Analysis,* 11: 93-108.

National Board for Professional Teaching Standards (1991). *Toward High and Rigorous Standards for the Teaching Profession* (3rd ed.). Detroit, MI: Author.

National Council of Teachers of Mathematics. (1989, March). *Curriculum and Evaluation Standards for School Mathematics.* Reston, VA: Author.

Quellmalz, E.S. (1991). "Developing Criteria for Performance Assessments: The Missing Link." *Applied Measurement in Education,* 4: 319-331.

Rebell, M.A. (1990). "Legal Issues Concerning Teacher Evaluation." In J. Millman and L. Darling-Hammond (eds). *The New Handbook of Teacher Evaluation: Assessing Elementary and Secondary School Teachers* (pp. 337-355). Newbury Park, CA: Sage.

Redfern, G.B. (1960). *How to Appraise Teaching Performance.* Columbus, OH: School Management Institute.

Redfern, G.B. (1980). *Evaluating Teachers and Administrators: A Performance Objectives Approach.* Boulder, CO: Westview Press.

Scriven, M. (1987). "Validity in Personnel Evaluation." *Journal of Personnel Evaluation in Education,* 1, 9-23.

Scriven, M. (1988a). "Duties-Based Teacher Evaluation." *Journal of Personnel Evaluation in Education,* 1: 319-334.

Scriven, M. (1988b). "Evaluating Teachers as Professionals: The Duties-Based Approach." In S.J. Stanley and W.J. Popham (eds). *Teacher Evaluation: Six Prescriptions for Success* (pp. 110-142). Alexandria, VA: Association for Supervision and Curriculum Development.

Sedlak, M., and Schlossman, S. (1986). *Who Will Teach? Historical Perspectives on the Changing Appeal of Teaching as a Profession.* (Report No. 3472—CSTP). Santa Monica, CA: RAND Center for the Study of the Teaching Profession.

Shalock, D., and Shalock, M. (1990, September). "Extending Teacher Assessment Beyond Knowledge and Skills: An Emerging Focus on Teacher Accomplishments." *Issues and Practices in Performance Assessment,* 2: 81-126.

Shimberg, B. (1985). "Testing for Licensure and Certification." *American Psychologist,* 36: 1138-1146.

Shulman, L.S. (1986) "Paradigms and Research Programs in the Study of Teaching: A Contemporary Perspective." In M.C. Wittrock (ed). *Handbook of Research on Teaching* (3rd ed., pp. 3-36). New York: Macmillan.

Shulman, L.S. (1987). "Knowledge and Teaching: Foundations of the New Reform." *Harvard Educational Review,* 57: 1-22.

Shulman, L.S. (1988). "A Union of Insufficiencies: Strategies for Teacher Assessment in a Period of Educational Reform." *Educational Leadership,* 46: 36-41.

Shulman, L.S. (1988a). "The Paradox of Teacher Assessment." In *New Directions for Teacher Assessment. Proceedings of the 1988 ETS Invitational Conference.* Princeton, NJ: Educational Testing Service.

Shulman, L.S. (1989). "The Paradox of Teacher Assessment." In *New Directions for Teacher Assessment: Proceedings of the 1988 ETS Invitational Conference* (pp. 13-27). Princeton, NJ: Educational Testing Service.

Stodolsky, S.S. (1988). *The Subject Matters.* Chicago: University of Chicago Press.

Streifer, P., and Iwanicki, E. (1987). "The Validation of Beginning Teacher Competencies in Connecticut." *Journal of Personnel Evaluation in Education,* 1: 33-55.

Webster, W.J., Mendro, R.L., and Almaguer, T.O. (1993). "Effectiveness Indices: The Major Component of an Equitable Accountability System." Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Wise, A.E., Darling-Hammond, L., McLaughlin, M.W., and Bernstein, H.T. (1984a). *Case Studies for Teacher Evaluation: A Study of Effective Practices.* Santa Monica, CA: Rand Corporation.

Wise, A.E., Darling-Hammond, L., McLaughlin, M.W., and Bernstein, H.T. (1984b). *Teacher Evaluation: A Study of Effective Practices.* Santa Monica, CA: Rand Corporation.

CHAPTER 8

EVALUATING TEACHER PERFORMANCE IN LATIN AMERICA

Franciso Álvarez Martín, in collaboration with María José Álvarez and Paula Vergara

The preceding chapter presented a broad panorama of teacher evaluation. This chapter addresses the same problem, but from the perspective of actual practice in a sample of five countries of Latin America. The document identifies a number of evaluation models proposed in the region by researchers or tested on a reduced scale through various educational innovations. In addition, it discusses the policy, technical, and resource problems facing each initiative designed to establish national systems that evaluate the teaching profession.

INTRODUCTION

This paper presents an overview of teacher performance evaluation practices in Latin America and makes suggestions for framing policy in this area based on Latin American Educational Information and Documentation Network (REDUC) data.

Of a total of 7929 works on education summarized as of the second half of 1996, we found only 78 titles (not all of which were equally pertinent) under the descriptor "teacher evaluation" in the collection published by REDUC. Just about all of these works have been published within the last ten years. Moreover, half the institutions sponsoring these studies are private institutions, 39 percent are universities, 6 percent are education ministries, and 5 percent are international organizations.

Most of these works (53 percent) are extremely general and, as such, do not refer to any one particular country or region. Another large portion of these works (27 percent) addresses the issue of teacher evaluation at a university level, from the perspective of university practices, grounded in a theoretical approach and in a study of practical experiences. The remaining works refer specifically to teacher evaluation in secondary schools or trade schools (13 percent), with very few (5 percent) approaching this issue at the basic education level.

The issue of teacher evaluation has not been a top priority in Latin America, which is not to say that there are no prevailing practices, legislation, or regulations in this area. Evaluations have always been conducted and are an ongoing enterprise in all educational institutions. Directors of educational institutions, as well as students and their families, are constantly evaluating the performance of corresponding teachers,.even if only with informal evaluation mechanisms. However, evaluation criteria and methods are far from standard. The only constant seems to be in teachers' reactions: they invariably regard any proposal for conducting systematic evaluations as a threat.

The educational reforms implemented throughout the region since the 1960s have focused mainly on expanding coverage and have had a technical emphasis, in which efficiency in education is regarded more a function of the quality or technical precision with which certain specific tasks are performed than of the attributes or qualities of the teacher in question.

These approaches helped achieve major breakthroughs in terms of providing universal access to education and improving annual school attendance rates, but have had no effect on educational dynamics within the classroom.

In the 1980s, the emphasis in education policy shifted to the promotion of educational effectiveness and quality, inspired by a series of studies conducted in preparation for the World Conference on Education for All (Jomtien, Thailand). Studies by Namo de Mello (1982) acknowledged the fact that quantitative educational development had imposed a need for major qualitative changes in educational dynamics. This research inspired a wave of studies of this issue throughout the literature, including studies by Ezpeleta (1989) and Tovar (1989) on the status of teachers in Argentina, Peru, and Bolivia, as well as studies of educational effectiveness conducted by Arancibia (1994) grounded in the analysis of factors affecting the quality of instruction.

This issue has taken on heightened importance in the past few years as a result of its ties with what appears to be the main emphasis of ongoing educational reform efforts throughout the area, as is made clear in the project presented by the National University of Colombia to the National Education Ministry for the implementation of a National Evaluation System (1996). The project paper underscores the fact that new educational concepts and objectives require teacher evaluations to motivate and prepare them to deal with the social effects of these reforms in a responsible fashion. Policies focused on maximizing the outreach of education systems have given way to new policies designed to guarantee higher-quality education; hence current concerns over what is being taught and what is being learned. While we know there are many different factors involved in what children learn, the practices of those in charge of the teaching and learning process are clearly a key factor and deserve to be given more emphasis.

Nevertheless, there is really no consensus around the need not only to study this issue, but to legislate it and promote programs and projects for the development of evaluation systems. While certain experts such as Connelly (1989) claim that teacher evaluations are time-consuming, overly costly, and ineffective at demonstrating benefits significant

enough to justify their continuation, Ahumada (1992), for example, maintains that evaluations of teaching efficiency are one of the most interesting and perhaps most overlooked issues in education processes and development efforts.

Depending on our position, either we need to realize that a lack of models and methods and an unfamiliarity with their use is an open invitation to improvisation and, hence, rejection by the subjects involved, as pointed out by Ahumada (1992)—or we should insist on the need to give priority to educating teachers-in-training rather than to providing inservice training for practicing teachers, as maintained by Connelly (1989).

The paper is divided into three parts. Part One looks at research data on this issue referring specifically to: a) the concept of teacher performance evaluation *per se*; b) evaluation criteria; c) achievements by innovative educational programs; d) evaluation models inspired by research; and e) the union position on teacher evaluation. Part Two examines existing policies, regulations, and practices in this area in selected countries. Part Three presents research and policy recommendations inspired by the studies and experiences examined in Parts One and Two.

WHAT DO WE KNOW ABOUT TEACHER EVALUATION?

The concept of "teacher performance evaluation"

Teacher evaluation systems and evaluation experiences throughout the region are variations on corporate evaluation systems. They were implemented in an endeavor to introduce into the school system the concept known in business administration circles as "organizational development," with a view to improving efficiency. However, while schools may want to be regarded as businesses, they will always have certain features associated with their goals and "outputs" that set them apart from other types of businesses. This fact may well be at the root of most of the questions, inconsistencies, and gaps confronted by teacher performance evaluation efforts.

Studies by Ahumada (1992) attempt to explain teacher criticisms and complaints in the face of any effort to evaluate their performance by the fact that, in most cases, the information is used for purposes whose end result is to threaten their job security rather than help improve their teaching.

Villa Sánchez (1985) maintains that, when we talk about teacher evaluation, we are not always referring to the same thing. The concept of teacher evaluation is different, depending on whether its purpose is occupational advancement or the improvement of teaching practices. Along these same lines, Connelly (1989) points out how teacher evaluation forms tend to include more superficial or routine administrative tasks than actual teaching-related tasks. Carranza (1992), in turn, underscores the fact that, rather than serving teachers by helping them improve their performance, the performance data is used as a coercive weapon. Thus, Villa Sánchez's assertion that every teacher evaluation system should include a definition of teaching tasks and a specific evaluation procedure should be interpreted as an effort to improve their coherence.

Moreover, certain experts feel that differences in current models also stem from the wide range of different instruments involved. While there is some consensus with respect to the main objective of teacher evaluation which, according to Carranza (1992), is to establish the educator's professional qualifications, preparation, and performance or, according to Garro (1988), to assess teacher performance, there is no single, technically sound instrument capable of supplying adequate information on the quality of a given teacher or the quality of that teacher's instruction. On the contrary, as pointed out by Ahumada (1992), we need to rely on several different instruments furnishing data from different sources, different points in time, or different procedures. This is the rationale behind the assertion by Ordoñez et al. (1996) to the effect that teacher evaluation should be considered a basis for the designing of educational strengthening plans and programs and, as such, requires making a series of examinations.

Nevertheless, we need to be aware of the fact that teacher performance evaluations are still in the developmental phase and, hence, require further research. At the same time, we must bear in mind that existing research in this area has given us certain incontrovertible general and specific indicators with respect to what most of us consider to be a successful teacher. This is certainly another important contribution by research in this area.

Evaluation criteria

It is by no means an easy task to establish precisely what "to evaluate" means as far as teacher performance is concerned. On one hand, it is hard to build a consensus around the profile of an effective teacher. Furthermore, we know that many factors other than teacher performance *per se* come into play as part of the teaching process, which, itself, is highly complex—hence the importance of research findings with respect to the types of criteria that need to be taken into account in evaluating teacher performance. These criteria have to do with teaching practices as well with as the effects and consequences of these practices.

The Avalos and Haddad study (1981) of teacher effectiveness endeavors to pinpoint teacher-related variables associated with changes in teaching conditions or in students, schools, and the community. The study divides these variables into two main categories, namely teacher attributes (sex, age, personality, socioeconomic status, knowledge, skills, language of instruction, attitudes, experience, qualifications, and training) and school system attributes (school location, type of school management, subject area, grade level, resources, examination system, work load, incentives, salary, and the teacher's social status). According to this study, most evaluations take into account the teacher's contextual background, attitudes, and classroom performance, but very few link these teacher-related factors to student achievement and changes in student attitudes.

Another approach, which has been in use since 1960, is the so-called "Descriptive Questionnaire on the Organizational Climate." The questionnaire helps establish the existence of an organizational reality or climate by assessing the performance of teachers

and school administrators. This technique, which emphasizes relational factors both in and outside the classroom, is a feedback instrument for helping teachers as well as their superiors decide on measures for overcoming weaknesses in organizational performance and, as a result, improve teacher performance and efficiency. However, according to Rossenfeld (1983), it is not easy to administer, and requires preparation on the part of the teaching staff and certain adaptations to the unique features of area schools.

A study by Galo de Lara (1990) on the evaluation of teacher performance in the education process in Guatemala focuses on teaching practices in the teaching/learning process. The study uses an observational instrument that lists a series of different categories and that can be adapted for use as a self-evaluation instrument. The different categories establish interrelationships that ensure the instrument's internal coherence while, at the same time, indicate which facets of the teacher's performance need to be improved. This predominantly qualitative instrument (see Table 1) is designed to inspire reflection and self-appraisal on the part of the teacher with respect to the administration of the educational process, so that performance can be improved.

The study regards evaluation as a *process* in which a series of variables come into play. The variables represent the building blocks of what actually takes place in the classroom, namely:

- teaching-learning activities
- teacher and student behavior
- interrelationships and types of interaction and the type of climate or atmosphere in the classroom.

According to the study, a good teacher is one who adheres to the normative elements of the model, as presented in the following table.

In a study of the attributes of teachers in Argentina, Justa Ezpeleta (1989) points out how the education system has always thought of teachers in abstract terms, without considering social factors or their personal aspirations or ambitions.

In point of fact, it is impossible to address the issue of teacher performance evaluations without taking into consideration the sociocultural context within which teachers develop their skills. Both in the study by Ezpeleta and in a similar study by Tovar (1989) of teachers in Peru, it is clear that everyday educational practices are affected by situations which have nothing to do with the teaching skills of the teacher alone. It is important to bear this in mind when trying to establish criteria for evaluating teacher performance.

Studies conducted by Arancibia (1987 and 1994) attempt to identify the practices of effective teachers (defining effective teachers as teachers whose students are high achievers), showing how student academic achievement is influenced not only by teachers' classroom practices, but also by their expectations, beliefs, feelings, and preferences.

Table 1. Categories and Scales for Observing the Educational Process

Category	Scale
1. Physical setting (The physical space and its fixed, mobile and semimobile elements, organized to promote learning)	1.1 Rigid-Flexible 1.2 Monotonous-Stimulating 1.3 Dirty-Clean 1.4 Disorderly-Orderly 1.5 Disorganized-Organized
2. Organization of time (Organization of the class schedule)	2.1 Rigid-Flexible 2.2 Structured-Unstructured
3. Use of time (Performance of teacher/student tasks in allotted class time)	3.1 Inefficient-Efficient
4. Organization of teaching tasks (Prior organization of tasks to be performed by the teacher)	4.1 Improvisation-Planning 4.2 Routine-Novelty 4.3 Ignoring the context-Taking the context into consideration 4.4 Separate subjects-Core subjects 4.5 Separate areas-Core areas
5. Types of teaching tasks (The teacher's designs in implementing the educational process)	5.1 Phases of the educational process 5.1.1 Motivation 5.1.2 Presentation 5.1.3 Development 5.1.4 Synthesis 5.1.5 Evaluation 5.1.6 Adjustment 5.2 Use of teaching aids 5.2.1 Oral-multimedia 5.3 Instruction-Practice 5.3.1 Oral-Oral and written
6. Organization of learning tasks (Organization and sequencing of student tasks)	6.1 Boredom-Interest 6.2 Insecurity-Security 6.3 Outside Control-Self-control 6.4 Rigidity-Flexibility 6.5 Individual work-Different types of group work
7. Types of learning tasks (Acts performed by students in order to learn)	7.1 Receptive 7.2 Reflexive 7.3 Reactive
8. Types of interaction (Primary role in communications)	8.1 Teacher as Transmitter-Student as Receiver 8.2 Student as Transmitter-Teacher as Receiver
9. Nonverbal communication (Nonverbal types of teacher behavior)	9.1 Distance-Rapprochement 9.2 Lack of control-Self-control 9.3 Indifference-Emotionalism
10. Emotional climate in the classroom (Confidence/security and classroom harmony between students and teacher)	10.1 Businesslike-Nonbusinesslike 10.2 Impersonal-Personal 10.3 Authoritarian-Democratic 10.4 Anarchical-Focused 10.5 Disagreeable-Agreeable

Ahumada (1992) maintains that "the incessant search for criteria, indicators, and procedures for evaluating teacher efficiency finally appears to be coming to an end," attempting to illustrate the extent of the consensus among researchers in this field in different countries. At the same time, Ahumada suggests that the time has come to use consistent criteria and, based on these criteria, to design versatile evaluation instruments in line with the objectives of corresponding evaluation processes. Inspired by the two-dimensional model developed by Popham (1980), which takes into account the way in which the data is collected and the objectives of corresponding evaluation processes, Ahumada adds a third dimension referring to the point in time at which the evaluation is conducted, thereby constructing a three-dimensional model grounded in standard criteria (see Table 2) as the basis for designing instruments attuned to the objectives of the corresponding evaluation process.

The 1989 study by Connelly is highly critical of the different approaches and/or criteria used to evaluate teacher performance: the author maintains that they invariably fail to take the teachers themselves into account. He points out how peer evaluation is a form of inservice education designed primarily to improve teaching. One of his misgivings is that, in certain developing countries, the level of professional expertise may not be high enough to justify this type of system. He suggests an alternative, which he refers to as a "supervised reflective practice," and which he views as a teacher education process rather than a system for firing or promoting teaching personnel. This process is a lesson in common sense on how teachers can reinforce their strengths by examining their teaching practices with supervisors and colleagues.

Table 2. Criteria for the Construction of Evaluation Instruments

1. *Adequate curriculum and classroom planning*, or the presentation of specific objectives and content, the methodological and evaluation strategy, and the recommended bibliography.
2. *Affective communication* (motivation—handling of materials), or the actual instruction process in which student achievement is improved as a result of the proper use of instructional aids, which can include written materials as well as factors such as the clarity of the message transmitted as part of the communication process.
3. *Knowledge of and enthusiasm for the subject area* (mastery of the subject), or the extent of the teacher's mastery of the field, as reflected in the ability to relate what is taught to problems encountered by students and associated with their environment.
4. *Positive attitudes towards students* (extent of the student-teacher relationship) is an important variable in the ongoing intercommunication process, reflected in the teacher's acceptance of student errors or different viewpoints, the teacher's accessibility in and outside the classroom, the teacher's continuous efforts to encourage student participation, an attitude of mutual respect on the part of the teacher and the students, and the teacher's willingness to give and receive advice.
5. *Flexibility of the educational process*, or the adaptation of existing resources and methods to student capabilities, the teacher's ability to allocate class time, work with small and large groups, select activities, and handle specialized materials according to the specific level of instruction.
6. *Impartiality in conducting evaluations and grading*, or the selection of criteria ensuring fair and impartial evaluations, as well as to the design, correction, and grading of tests based on sound technical principles ensuring their authenticity and reliability.

Connelly's so-called "supervised reflective practice" is another version of what Shön (1987) refers to as "reflection on action," which is an intentional endeavor to reassess one's past actions and pinpoint those things that could have been done better. In its simplest form, this is a three-stage process consisting of

- an observation session
- a discussion session on the findings from the observation session
- a follow-up discussion, possibly triggering further observations

This three-tiered structure can be expanded and modified in different ways. It can be used just as effectively in the training of student teachers as in inservice training activities for practicing teachers.

Research findings as well as existing practices can help clarify certain concepts. In fact, there are a number of innovative experiences propounding new types of approaches and procedures. An example of one such innovative concept is the notion that teacher evaluation should be a participatory process involving the subject, rather than a procedure administered solely by outside agents. Along these same lines, Ordóñez et al. (1996) argue that evaluations can help reform education only if conducted with the participation of stakeholders directly in charge of the everyday operation of the school in question.

Achievements by innovative educational programs

Successful experiments in education that can be categorized as innovative in the sense of implementing new strategies with qualitative effects on student learning have underscored the importance of establishing participatory mechanisms enabling teachers to review, critique, and modify their performance with a view to making qualitative improvements in their practices. For purposes of this paper, these experiences are discussed exclusively from the standpoint of their teacher performance evaluation components.

Colombia's "New School" program

Teacher evaluation at the "New School," an innovative educational program launched in Colombia (see Schiefembein, 1993), is conceived of as an ongoing examination of educational practices, both for decision making purposes and for making qualitative improvements in such practices. This examination process basically involves teaching personnel and other interested parties such as microcenter coordinators and supervisors. However, it also includes third parties outside the educational institution, such as experts and researchers engaged primarily in studying performance and corresponding determining factors.

a) A network of microcenters provides teachers with a regular, organized forum for interacting with their colleagues and sharing their teaching experiences. This allows them to examine problems and discuss the effects of their teaching practices, which

helps allay teacher uncertainties and fears associated with the use of new teaching methods. It also affords teachers an opportunity to critique their own practices, discuss solutions to emerging problems, and seek and establish new objectives. This means of examining their actions helps point teachers towards specific, innovative solutions and helps create a frame of reference for the reorganization of efforts at the school level.

b) Local microcenter coordinators and supervisors provide guidance at monthly meetings, which are also, by nature, a kind of monitoring process—they follow up on the implementation of innovations, which transforms these meetings into a source of valuable evaluation data. This guidance and monitoring mechanism was declared an official New School supervision process in 1985. Each supervisor must present an evaluation report on the performance of the microcenter in question and of all participating teachers.

c) Another important element of teacher evaluation at Colombia's New School is rooted in empirical research on student achievement and the determining factors involved. Tracer studies as well as expert opinions claim improvements in promotion rates, test scores, self-esteem, and teacher satisfaction. There is improvement as well in the support forthcoming from local officials and the community at large, as alleged in a UNESCO back-to-office report on an observation and evaluation mission conducted in 1985, an evaluation report by Rojas and Castillo at the SER Research Institute (1987), and a 1992 evaluation report by the World Bank, in which it rates Colombia's New School as one of the top three experiments in primary education.

d) Lastly, commentaries by a number of experts, who visited these schools personally and were able to observe teacher performance in the classroom firsthand and discuss their observations with the teaching staff, are another source of valuable information for teacher evaluation purposes.

Chile's 900 Schools (P-900) program

With the transfer of power to the country's new democratic coalition government elected in 1991, the Chilean Ministry of Education focused its education policy on improving the quality of education, promoting equity by expanding education opportunities for the poor, and getting everyone involved in the education process. In so doing, Chile mounted a new educational program known as the P-900 program targeted at the country's 900 poorest schools. Program components included infrastructure improvements; the supply of textbooks, classroom libraries and teaching materials; and the conveyance of inservice teacher training workshops and student learning workshops (known as TAPs) with community participation.

The program also included a teacher performance evaluation component as an integral part of its inservice teacher training workshops which, initially, were held weekly (and later, biweekly), at the individual school level. These teacher-training workshops were conducted by supervisors at each school. The goal was to juxtapose thought and reflection on prevailing teaching practices with the introduction of new teaching

methods. In addition to addressing specific subject areas, these meetings generally helped promote the sharing of experiences and were used by the teachers as a way to ease the pressures of dealing with the various problems confronted by their students.

This created a forum at individual schools for peer discussions of educational issues such as student achievement and teaching practices, prompting teachers to reflect on their actions, evaluate their performance, and look for ways to improve their performance (García-Huidobro, 1994). Such forums empower teachers in the practice of their profession, allowing them to be innovative and creative in their teaching practices and holding them professionally accountable for their performance.

As pointed out by Filp (1994), this activity embodies one of the program's guiding principles for improving the professional status of teachers, namely that of helping teachers make curriculum-related decisions based on their own judgment while, at the same time, suggesting changes in their practices rather than attempting to impose such changes upon them as ready-made formulas. In this sense, it can be considered a teacher performance evaluation process encompassing classroom teaching practices *per se*, as well as other supporting administrative tasks, in which there is no perceivable opposition on the part of corresponding teachers.

Moreover, according to this same study (Filp, 1994), this evaluation component also affects the performance of corresponding supervisors who, from what was essentially a regulatory function come to play a more technical, supervisory role (through their empowerment by the training dispensed to them under the program's instructional component). This fact is recognized by the overwhelming majority of supervisory personnel (97.5 percent of the respondents in a corresponding survey), who claim that their participation in this program enriched their professional performance.

In an external evaluation and monitoring report on this program, Filp (1994) underscores the importance of this self-evaluation component in the following terms:

> Self-evaluation also plays an important role in the regular monitoring of the quality of work performed. Thus, each team of supervisors conducts a yearly province-wide performance evaluation, which is discussed at regional meetings attended by heads of provincial government departments and program coordinators. A package developed in 1991 allows the teaching staff of each school to conduct a participatory evaluation of its educational performance at the end of each school year.

Models inspired by research

There are many different approaches to evaluating teacher efficiency, depending on the goals of the evaluation in question. Thus, teachers can be evaluated for administrative purposes such as promotion, compensation, or occupational advancement, or for educational purposes where the goal is to improve the quality of their teaching. This

wide variety of models is a product of the different theories of learning advocated by prevailing systems and organizations.

The purpose of examining different research-inspired models for evaluating teacher efficiency is to provide a frame of reference to help us better understand teacher performance evaluation practices in certain countries that are in the process of reforming their education systems.

Teacher profile-based model

According to this model, teacher performance is evaluated from the standpoint of its consistency with the traits and attributes of what is considered to be an ideal teacher, based on a pre-established teacher profile.

These attributes can be established in one of two ways. The first is to develop a profile of the perceptions of different groups (students, parents, colleagues) as to what makes a good teacher. Since we've all been to school, we all have an opinion about the attributes of those teachers we remember as being good at their jobs. The perceptions of these different groups (students, parents, and teachers themselves) are used as building blocks to establish the attributes of the so-called "ideal teacher." Obviously, this profile is going to include certain personality traits, as well as attributes associated with technical/educational factors. Authoritative studies such as those of Charters and Waples find the top-ranking attributes to be versatility, consideration, enthusiasm, good judgment, honesty, and charisma. More recent studies in this area, such as that of González Soler (1980), confirm that the most effective teachers are those who appear to be more human in the broad sense of the term or, in other words, pleasant, affectionate, fair, democratic, and better able to relate to their students.

Another way to establish the profile of a good teacher is through firsthand and indirect observation, pinpointing essential teacher attributes associated with student achievement.

Once this profile has been established, the next step is to design questionnaires that can be administered in different fashions, namely as a means of self-evaluation—by an outside evaluator in an interview with the teacher in question, or through a survey of corresponding students.

This model has drawn both praise and criticism. One of its positive features is that teachers are identified after the instruction is dispensed, the inference being that teachers who make the best impression are efficient and can be identified as such by their students. On the other hand, one of the model's negative features is that in conducting corresponding observations, little is known about the long-term effects of the teacher's performance, because this evaluation method is grounded in an ongoing education process. Moreover, few studies of teacher attributes distinguish between basic human traits—which are influenced little by teacher education processes—and educable factors, such as mastery of a curriculum content area or an ability to ask questions. The most problematic aspect of this model is that it is based on the profile of a nonexistent

teacher whose attributes are virtually impossible to implant in future teachers, since many involve traits that cannot be easily taught in training programs.

Other criticisms of this model have to do with the difficulty of reaching agreement as to whether a given subject is a good teacher, and the lack of a strong correlation between teacher attributes and student achievement. According to its opponents, this model fails to produce an objective, reliable evaluation of a teacher's attributes and capabilities. As maintained by Connelly (1990), judging teacher effectiveness on the basis of the teacher's visible traits is risky at best.

Moreover, while one of the criticisms of this approach is that corresponding indicators are based on the perceptions of students, who are regarded as incapable of evaluating certain types of factors, other commentators such as Ahumada (1992) maintain that there are certain areas in which the opinions of students are essential and can furnish sound, reliable information.

This model has inspired the development of interesting evaluation instruments and has helped clarify certain concepts. At the same time, its inability to reconcile criteria or standards set from different perspectives, namely that of the teacher and the student, illustrates the relativity of the concept of the so-called "good teacher."

Student achievement-based model

The main feature of this model is that it evaluates teacher performance by verifying the learning or achievements of corresponding students. It is inspired by a current of thought that is highly critical of the school system and of what is being done in the school. Advocates of this model maintain that the proper criterion for evaluating teacher performance is not one in which the emphasis is on the teacher's actions, but rather on their effects on students. Using this criterion as the basis for gathering information for teacher evaluation purposes poses the risk of compromising the quality of education in that, as pointed out by Ahumada (1992), "knowing that they are being evaluated based on the achievement of their students, teachers tend to focus on teaching lower-level replicable processes, neglecting higher-level processes that are inherently more difficult for students to grasp." Moreover, such an approach is unsound in that teachers cannot be held entirely accountable for the success or achievement of their students. The study by Cardemil (1991) has a great deal to say on this matter, maintaining that student achievement is affected by many different factors, only one of which is the teacher. In a sense, this model undermines the teacher's work in that it fails to recognize the complexity of the teaching process and narrows the concept of learning to the mere transmission and replication of facts.

Classroom performance-based model

In this model, the emphasis is on teaching as a process and its correlation with output variables such as student achievement. Its premise is that teacher efficiency is best evaluated by identifying those aspects of the teacher's performance that affect student

achievement, and that relate to the teacher's ability to create a propitious learning environment in the classroom.

This has been the model of choice since as far back as the 1960s, employing observation techniques, input-output tables, or different scales for assessing teacher performance.

Criticisms of this evaluation model are leveled primarily at the party in charge of conducting the evaluation. The main objection to this model is that reported data reflect observers' personal conceptions of the elements of effective teaching and are confirmed by their own standards for each act observed. Subjectivity is definitely an issue and allows observers to reward or penalize the subjects under evaluation for reasons totally unrelated to teaching efficiency and having more to do with feelings of sympathy or animosity toward the subject.

Moreover, Ahumada (1992) maintains that a prior knowledge of the criteria, evidence, and standards used in corresponding observation procedures can prompt teachers to make certain changes in their classroom teaching practices. In some cases, teachers fail to make proper use of personal qualities conducive to promoting interaction and, thereby, reduce their effectiveness as teachers. The use of this model can help produce interesting observation techniques and instruments, provided they are based on empirical observation and grounded in each country's own specific approaches to education.

Reflexive practice model

This model embodies supervised reflection. The objective of evaluation in this case is the strengthening of educational personnel—rather than supervision or monitoring as the basis for layoffs or promotions. In this sense, the model supplies information for decision making aimed at improving the quality of education. The model conceives of teaching as a sequence of problem-targeting and problem-solving events in which teachers are continuously developing their skills as they confront, define, and solve practical problems through what Schön (1987) refers to as "reflection on action." This practice requires that teachers reflect on their actions or evaluate them "after the fact" to examine their accomplishments and their failures and consider what types of things they might have done differently. The implementation of this model is a three-phase or three-stage process, including

- an observation session and anecdotal report on the class in question;
- a reflective discussion between the subject and the observer to comment on corresponding observations and ask questions designed to expose the relevance and coherence of the practices observed; and
- a follow-up discussion to review the issues previously addressed and the measures agreed on in the second phase of the process. If necessary and advisable, this phase of the process can also include a second observation and report.

The use of this model requires a supervision system staffed by supervisors who have allotted sufficient time for this purpose. However, the model can be adapted to allow

corresponding observations to be made by other parties, such as by colleagues within the same school facility or by a teacher in an administrative position.

As a model with built-in versatility, a strong participatory element, and an educational focus, in the sense that resulting information is used to strengthen or improve teaching practices, it is equally well-suited for evaluating student teachers.

Teacher evaluation and the teachers' union

As far as teachers' unions are concerned, there are two main trends in Latin American nations, namely the promotion of a process of "proletarization" and the reinforcement of the professional and technical nature of teaching.
At the same time, Latin American teachers' unions tend to be highly politicized due to conditions unique to that area, such as a strong central government presence in educational management and service delivery, which makes government an employer as well as a legislator in this area, and by political parties' role as intermediaries in dealings with public authorities (due to the relative weakness of union movements). This situation is further aggravated by a strong sense of grievance triggered by teachers' employment conditions. According to the study by Nuñez (1990), this situation has created a tendency at the central government level to exclude union organizations from decision making processes affecting education policy. Union movements within the teaching profession have focused more on institutional and administrative issues than on the school curriculum or teaching itself.

Union organizations have not taken advantage of educational research findings, because they are suspicious of their allegedly technocratic nature. Educational reform efforts have made no provisions for promoting creative participation by union organizations. Moreover, as pointed out by Nuñez, in many cases, union reactions to these reforms have been hostile. There are few cases of educational innovations where existing union organizations were not reactive, did not involve haggling or an outright rejection of the innovation, or came up with an alternative proposal.

One of the few exceptions to the rule was the Chilean National Teachers' Association which, during the period from 1923 to 1927, was continually critical of the education system, eventually framing its own organizational and educational project. The criticism triggered an important, if ephemeral, sweeping reform of the country's education system in 1928. In another instance, the Colombian Federation of Educators (FECODE) made one of the most interesting and innovative contributions to educational development inspired by a union movement. Mejía (1987) explains how, in addition to promoting change within the union itself, the "education movement" (springing up within the ranks of union organizations in this country) "set in motion a process of self-examination calling for teachers to critique their own performance without hiding behind the union mantle of protection, requiring a level of professionalism on the part of the teacher giving him a sense of professional identity . . . allowing him to reflect on his background and retrace his professional experience as a way to identify with his actions."

The Colombian education movement encouraged teachers to cease being mere parrots and to become builders of a new society. This meant taking the lead in mounting an evaluation effort that as an integral part of their everyday work, would not generate internal opposition, as it did not refer to administrative or organizational facets of their work, but to building a store of knowledge for society. According to Mejía (1987), those involved in the education movement were fully aware that this involved a learning process. By having a new voice in education, teachers needed to be prepared for this role by learning to speak out on their work, overcoming the fear that this inspired, such as the fear of speaking out, or of being supplanted by education "experts." Above all, this learning meant accepting evaluation as a necessary part of their responsibilities, as recognized in the guiding principles of the education movement and, more specifically, in the stipulation that teachers should reflect on their practices, discuss them, and embark upon an ongoing process of evaluation and self-examination.

Education movements foster reflection and allow teachers to move from a purely technical into a more professional role and helps them discover for themselves the reasons why they teach.

Other education movements include the Simiente group in Brazil, the Antonio Encinas Schools in Peru, and the Espacio and Freinet groups in Chile.

TEACHER EVALUATION PRACTICES

The following experiences with teacher evaluation were compiled by the REDUC network. There may be other such experiences; however, there is no written documentation available from this network. In any event, the cases examined in this paper represent a wide range of trends, circumstances, and aspirations to help us better study this issue and, perhaps, suggest areas for further work.

Argentina (San Juan Province)

Law 2492 passed in 1976, known as the Teacher's Act (Chapter XV, Articles 82 through 85), specifically requires the evaluation of teacher performance, which is to be tracked in a personal performance record and which the teacher has the right to inspect and verify.

These evaluations are conducted annually by a board of examiners and cover such areas as the teacher's general cultural and professional background, teaching and disciplinary skills, and attendance and punctuality. There is a standard form for recording the results of the professional assessment. The law also establishes appeal procedures for contesting assessments. Repeated attempts to amend this statute have not been successful.

Despite these regulations, evaluation practices appear to be deviating from their intended purpose and, as far as teachers are concerned, do nothing to improve teacher performance. Moreover, they produce innumerable disputes. Some are settled either by giving all teachers top performance ratings, thereby turning the evaluation process into

a useless ritual, or allowing long, tedious appeals when there is no basis to support a poor performance rating.

Under the provisions of the law, these teacher evaluations are the responsibility of the entire team of school administrators in each educational institution. However, in practice, the task is left up to the school principal who, in turn, "passes the buck" to the board of examiners, with all the risks this entails.[1] There is widespread belief among teachers that the reporting form used for these yearly evaluations is useless. Moreover, when supervisors, in an attempt to improve monitoring, seek to gather more information on how classes are being conducted, they are informed that school principals have no time for this activity.

Experts on the faculty of the National University of San Juan and its various colleges are working to change teacher evaluation policy. One of the recommendations made by advocates of educational innovation calls for educational institutions to establish teaching teams of teachers of courses in related areas and to train them to observe each others' classes.

The teachers' union movement in Argentina, with its broad base and diverse opinions, does not appear to have an official position on evaluation, as its main concern is the issue of pay. There is currently strong union opposition to any government plan to begin evaluating teachers from the standpoint of their productivity, arguing that it is wrong to base teacher evaluations on indicators such as how many students pass their courses or on how many students attend class.

Colombia

The General Education Act (Law 115 of February 8, 1994) provides for the establishment of a National Educational Assessment System to ensure educational quality, the attainment of educational objectives, and better teacher training. According to this legislation, the purpose of the envisaged system is to design and implement evaluation criteria for assessing the quality of education; the performance of teachers and school administrators; the achievement of students; the effectiveness of teaching methods, textbooks, and instructional materials; and the structure of administration. It establishes two evaluation mechanisms, namely yearly performance evaluations at the school level, and a mechanism for testing the academic qualifications of corresponding educators within their specialized teaching fields as well as their familiarity with new developments in professional teaching practices at six-year intervals.

The fact that teachers as well as school administrators who fail this test are given a second chance is noteworthy. Teachers can retake the examination within one year. School administrators also have a year in which to submit a proposal for resolving any existing problems.

The Education Act calls for this system to operate in conjunction with the National Testing Service run by the Colombian Institute for the Advancement of Higher Educa-

tion (ICFES). The Act also sets rules for hiring new teachers and for their attachment to the civil service; it also contains provisions with respect to the status of currently employed teachers' job tenure. The ICFES and other organizations have begun formulating proposals and preparing projects in line with the new legal framework for the implementation of its provisions.

Under the project for the evaluation of teachers in service designed by the National University of Colombia (1996), evaluation is conceived of as a process. Examination is a specific point at which an assessment is made for purposes of framing strategies designed to improve educational quality. According to this project, evaluations of teacher performance should take into account all facets of the teacher's educational performance rather than only certain aspects of the teacher's work. In other words, evaluations of teacher performance should be conceived of as the basis for framing plans and programs for educational advancement, and not as a screening strategy for the eventual adoption of exclusionary or punitive measures. Accordingly, evaluation should be an integral part of the complex realm of educational practice and conceived of both as a process of compiling information from different sources and as an educational skill with

- a **communications dimension**, referring to the flow of the conceptual and informational content of educational interactions and the languages making this possible;
- an **ethical dimension**, referring to behavioral models facilitating the establishment of an appropriate set of values for a democratic society corresponding to an academic culture;
- an **aesthetic dimension**, having to do with the teacher's own sense of discovery and the manner in which students are guided through their discovery and learning process, with the only way to study this dimension by examining corresponding educational strategies; and
- a **psychosocial dimension**, referring to the teacher's ability to give students what they need, without overlooking what motivates them.

While teachers' reflections on their own teaching practices are essential, this process is greatly enriched by a comparison with the practices of other teachers—hence the importance of self-examination and peer evaluation. Student opinions are another important source of information on the teacher/student relationship.

The criteria that serve as the framework for the evaluation system and for corresponding evaluation instruments will refer to:

- The teacher's ability to relate the school environment to the outside environment. This requires the adaptation of scientific and cultural knowledge to make it relevant and understandable, both within the school environment and in the everyday life of the surrounding community. It also involves the screening, ranking, reconstruction, and reorganization of this knowledge.
- Mastery of educational languages. In other words, the teacher must be able to interpret written textbooks and various codes, as well as the language and drawings of

students. Teachers need to interact with their students and to promote a communications ethic, which implies a commitment to the written and spoken word.

- Teachers' inventive capacity. On one hand, the teacher must demonstrate an ability to adapt textbooks and school curricula to specific classroom conditions and situations and, on the other hand, must understand that there are certain conditions that need to be present to encourage innovation on the part of the teacher.
- Social commitment. Schools are inevitably socialization media, which vests the teacher with a social responsibility that requires an understanding of the implications of actions, as well as of limitations imposed by society itself or by the community.

The evaluation system is designed to help judge teacher capabilities from several angles or perspectives through the following components, based on the criteria outlined above:

Self-evaluation. The goal of this process is to inspire thought and reflection on the part of the teacher with respect to performance, conditions strengthening or hampering performance, and impact on society and the community. The process should cause teachers to examine factors such as their education and refresher training; the quality of their teaching and of school projects, textbooks, and curricula; the quality of relations with students; problems (infrastructure, training, economic or cultural problems) hampering their work; and relations with peers, school administrators, and the community.

Testing of teaching skills and the teacher's mastery of educational "grammars." These judge the ability to adapt knowledge, methods, and strategies to specific working conditions and to master basic theory within the field of specialty.

The concept of "general educational grammars" refers to every teacher's ability (regardless of specialty or the course being taught) to interpret written textbooks and other codes (mathematics, graphics, diagrams, etc.); to organize work space; to perform systematic observations; and to conduct different types of assessments. The term "specific educational grammars" refers to the teacher's ability to comprehend and implement basic principles, methods, and strategies associated with each subject area.

Open-ended question. This is a short essay-type question on a teaching-related subject designed to force the teacher to utilize knowledge of the theory and practice of teaching, as well as reasoning skills. The essay is graded by teaching personnel based on criteria such as the internal coherence of the text, the strength of corresponding arguments, the richness of the author's experience, the author's ability to express thoughts, and the author's writing skills.

Peer evaluation. The aim of peer evaluation is to examine the ability of teaching personnel to work together effectively. This evaluation is conducted by a collegiate body.

Student evaluation. This type of evaluation is designed to assess how the teacher interacts and communicates with students, as well as the teacher's work commitment.

Survey. This process looks at availability and the use of resources, relations between the school and the surrounding community, and the emphasis of subject matter and teaching methods. The survey put the teaching process in its proper context for educational mapping purposes and for different areas and contexts.

Nine standard tests, including one standard test for the basic primary education level and eight standard tests for the basic secondary education level focusing on different specialties, were developed in order to administer this performance-testing instrument.

The test is divided into three parts: educational concepts (approximately fifty test questions); general educational "grammars" (this section is the same for all the standard tests and consists of approximately thirty test questions); and specific educational "grammars" in each specialized teaching area (this section is different for each of the nine standard tests for basic education teachers and consists of approximately thirty test questions).

The maximum test score is equal to the average score obtained by the top 20 percent of highest-scoring teachers. The minimum score is equal to 20 percent of the maximum score.

Pending the issuance of regulations under the General Education Act governing the screening system for the hiring of teaching personnel, current practices are in keeping with previously established procedures and with the provisions of a 1989 decision establishing guidelines, formats, and procedures for the administration of competitive examinations conducted at the district and departmental levels. Technical, administrative, and educational assistance is available for this purpose from the Teacher Evaluation and Competitive Examination Unit attached to the Ministry of Education.

The competitive hiring process includes:

- a test of the applicant's general cultural background, teaching and administrative skills, academic knowledge, and educational expertise, by subject area. This is a screening or elimination-type test and represents 60 percent of the applicant's total score.
- an interview, which is also a screening or eliminatory process, representing 20 percent of the applicant's total score;
- an examination of the applicant's curriculum vitae, which is scored as follows: 5 percent for applicants born in the locality in question; 5 percent for five or more years of experience; 10 percent for experience in a rural area.

Costa Rica

In Costa Rica, teacher performance evaluations are given high priority as a means of verifying the contribution of its teachers to the achievement of corresponding educational objectives, as evidenced by the numerous studies in this area, the most noteworthy of which are outlined below.

- The psychoprofessional profile designed by the Ministry of Public Education's Personnel Department in 1969 establishing the personal attributes and professional qualifications of teaching personnel and corresponding evaluation criteria.
- The study launched in 1974 by the Multinational Center for Educational Research (CEMIE) in conjunction with the Educational Research Unit attached to the Ministry of Public Education on the validity of the performance evaluation system for teaching personnel, helping to establish and define desirable attributes for Costa Rican teachers.
- The quality control plan launched by the Educational Research Department attached to the Ministry of Public Education beginning in 1983, which included the conduct of studies for the construction of three profiles of ideal urban, multi-grade, and rural schoolteachers. These studies were followed by a second series of studies designed to establish necessary personal attributes and professional qualifications for natural science, social science, and language teachers.

The earliest information available on teacher evaluation in Costa Rica is found in Chapter V of the Education Code enacted back in 1944, referring to a performance record for keeping track of teacher performance throughout the course of the school year, to be used for promotion, pension, and other purposes. According to the Code's provisions, this performance record was to be filled out by the principal of the educational institution at the end of each academic year, and was to contain information on the teacher's work schedule, punctuality, quality of work, and on any incentives or reprimands bestowed. This same legislation also established an Examining Board for Teaching Personnel, headed by the Technical Director of Primary Education and consisting of regular Board members. Its responsibilities included rating teaching personnel, approving promotions, and establishing a general teacher roster.

The Basic Education Act of 1957 defined the objectives of education in Costa Rica, the organizational structure of the nation's education system, and the different stages or levels of education. The law charges the Ministry of Education with full responsibility for drafting legislation and corresponding implementing regulations. It calls for the enactment of a Teachers' Act grounded in democratic principles of public education, providing, among other things, for evaluations of teacher performance.[2]

In 1969, in line with recommendations presented by the committee drafting the organic act creating the Ministry of Public Education, the Ministry's Personnel Department published a handbook calling for the implementation of a performance evaluation and rating system for teaching professionals.

The resulting system included two basic elements: a standard form for evaluating all teaching personnel, and a handbook for the evaluation and rating of professional services rendered by teaching personnel.

However, evaluation system personnel and evaluees alike were dissatisfied with the evaluation process and with the evaluation instruments. The findings from a study by Fallas, Herrera, Páez, and Zamora (1993) of the civil service rating system for teaching

personnel revealed a pervasive tendency on the part of observers to overrate their subjects. The frequency of excellent ratings was clearly higher than would be expected in a normal distribution. School principals dared not disqualify anybody, except in obvious cases. Moreover, there was also evidence of a halo effect, in which the evaluator forms a general impression of evaluees and rates them in each specific area based on an overall impression—without making a separate, independent evaluation of each facet of performance.

In 1979, a report by the National Planning and Programming Committee claimed that the system presented a number of problems and called for a review of corresponding procedures. The performance evaluation and rating form was revised again in 1983. Education experts and union organizations criticized the changes introduced in this form as being inadequate and more superficial than substantive. In 1983, the size of the form was reduced and a series of boxes was added for the evaluator to rate the subject in each specific area. A more complete set of instructions for filling out the form and a new, detailed description of each evaluation area were added in 1985. However, teacher associations were anxious to build a consensus with the Education Ministry around the need to update the Teachers' Act and pushed ahead with a project for reforming the performance evaluation and rating system, in which proposed changes were, again, mostly a matter of form.

The study by Sánchez (1992) presents the results of a teacher survey conducted in connection with the proposal of a new performance evaluation and rating model for teaching personnel. According to the survey data, the rating scale of excellent, very good, good, inadequate, and unacceptable appeared to be widely accepted (by 83.3 percent of the teachers surveyed). Likewise, 64 percent of the respondents approved of the concept of self-evaluation. In general, the teachers agreed that the set of variables covered by the evaluation instrument should be kept intact. In the judgment of some teachers, it wasn't the performance rating instrument itself that was causing the problem, but rather the way it was being used. Paradoxically, 80 percent of the respondents felt that, for the most part, it was administered impartially. Moreover, most respondents were unable to recall the objective of the performance evaluation and rating system as defined in the Teachers' Act, the purpose of which was to bestow incentives and benefits such as transfers, promotions, educational leave, yearly pay hikes, grants, and use of various facilities.

Chile

Everyone is talking about Chile's achievements in promoting educational development and in terms of the social impact of its educational reform programs. This would seem to point to the existence of successful experiences with teacher performance evaluation processes as well. As in our previous discussions, it might be best to begin with a look at the different factors at play in this area.

One such factor involves changes in regulations governing the employment of teachers during the past twenty years. According to Núñez (1996), teachers in Chile no longer

fall under traditional civil service regulations. Rather, they are governed by new decentralized personnel management regulations that try to strike a balance between the various national government regulations designed to guarantee equity and respect for the legitimate rights of professional educators. Teacher performance evaluation was relegated to a secondary plane during the course of this transition period, with both the government and the teachers' union more concerned with improving working conditions than with framing policy in this area.

As of 1980, the country's public school teachers were no longer regarded as civil servants, but rather as municipal government employees with employment contracts, like any other worker governed by general labor legislation. As a result, teaching personnel lost certain rights, such as tenure which, in turn, began creating large salary differentials.

This, in turn, gave rise to a powerful union movement seeking improvements in employment conditions. In 1990, with the country's return to democracy, the union movement gathered new strength. It pushed for a formal statute for teaching professionals, which eventually produced tangible results, with the enactment of a Professional Educators' Act in 1991, which:

- placed teachers hired at all levels of the nation's public and private school systems and under all types of arrangements within a single legislative framework;
- tangibly improved job security for teaching personnel by automatically granting tenure to all teaching professionals holding contracts for indefinite terms as of 1991;
- allowed for transfers and teacher exchanges within the same municipality without any loss of seniority rights, and stipends for advanced training, as had previously been the case;
- regulated the length of the workday, set maximum work loads, determined the assignment of workloads, legalized summer vacations, and established the right of teachers with more than thirty years of service to be assigned lighter work loads without a corresponding cut in pay;
- rather than discriminating against teachers in private schools, afforded them unprecedented opportunities for equalization.

Discussions about the implementation of this statute focused on issues relating to the improvement of employment conditions for teaching personnel, neglecting other matters more concerned with qualitative improvements, such as evaluations of teacher performance, with respect to which there was a long tradition of opposition. The implementing regulations issued under this statute contain the following language:

> Professional educators are personally accountable for the performance of their respective teaching responsibilities. Accordingly, they are to be subject to performance evaluation processes and to be informed of the results of these evaluations. Professional educators shall have the right to appeal any direct assessment or evaluation of their performance that they consider to be groundless. The purpose of performance evaluation

> is to assess the performance of professional educators for all allowable purposes and effects under the present regulations.

In a paper presented to the government, the Technical Advisory Committee for the National Talks on Modernizing Education (1994) underscored the need to pursue ongoing efforts to improve employment conditions and further strengthen the professional status of teaching personnel. In other words, part of the same effort to improve the professional status of teachers, the paper called for the implementation of mechanisms that would create appropriate pay incentives, recognize personal initiative, and reward good job performance and resulting achievements. It refers specifically to so-called "merit pay" systems instituted in other countries.

Throughout this period, the country's teaching professionals, backed by the Professional Teachers' Association, stubbornly opposed any attempt to evaluate teacher performance. Their objection to the evaluation process recommended by this statute was that it was grounded in a punitive type of performance rating system, evaluating teaching personnel based on factors other than their quality as teachers. Thus, they rejected the system propounded by the statute and sought to negotiate another type of arrangement.

A joint committee of Education Ministry and Teachers' Association representatives is currently considering a new proposal for an evaluation system. According to the Teachers' Association, the proposal currently under consideration includes important innovations such as separate evaluations of different categories of educational personnel (school principals and teacher/administrators, technical/education unit staff members, classroom teachers).

The proposed system is designed to evaluate attitude and performance within the context of the surrounding environment, putting a premium on leadership and extracurricular activities. For example, peer evaluations would assess activities mounted within the school community. It is important that school principals not simply evaluate, but gather information and check it out before actually conducting the evaluation. The proposal also calls for the maintenance of a performance or service record on each teacher, whose contents would be divulged to the teacher in question.

The purpose of this evaluation process is to rate teaching professionals in five areas, or on five items, as well as to establish a merit list (which would be useful in cases of teachers competing for positions, requesting grants, or seeking incentive bonuses) and a demerit list (to be used for improving educational quality; teachers who appear on this list two years in a row would be disqualified from teaching in the public education system). According to the union, this project should already be in the implementation phase.

Opposition to performance evaluations within the ranks of the teaching profession is clearly illustrated in studies of teacher perceptions of supervision which, according to the Technical Teacher Supervision Handbook published by the Ministry of Education in 1990, performs an evaluation function. The study by Rubilar and Cuevas (1996)

reveals the extent of teacher dissatisfaction with the supervision provided by outside supervisors or supervisors attached to education departments at the provincial level. The general consensus among teachers is that, in addition to their infrequent school visits, these supervisors furnish them with no relevant guidance or suggestions and fail to motivate them. On the other hand, they value the supervisory work performed by senior technical personnel or internal supervisors within their respective schools, with whom they are in regular contact and from whom they welcome suggestions.

Moreover, in their eagerness to implement the new education policy framed by the country's new democratic government (whose main objectives are to improve educational quality, achieve equity by equalizing educational opportunities, and heighten participation by different social groups in the education process), national education officials have mounted a number of projects and programs that successfully promote innovation and strengthen the professional status of teaching personnel by including performance evaluation among the necessary elements for improving teacher performance.

The educational advancement projects (PMEs) are a case in point. These projects have been designed by teachers at the individual school level based on what they felt was important and needed to be done in order to reform the education process and improve student learning. This very process requires that teachers examine and judge their own teaching practices and look for alternatives to help improve student achievement. It also allows teachers a greater sense of professional satisfaction at the prospect of mounting projects that they personally designed. A total of 3,111 PMEs were conducted during the period between 1992 and 1995, with more than 75,000 teachers involved in their preparation. According to the Directory of Educational Advancement Projects published by the Ministry of Education (1996), as of 1995, there were 1,680 schools at the basic education level with ongoing PMEs.

There are other programs with similar approaches that also include an evaluation component as a way to help modernize teaching practices and raise the quality of education. However, these sorts of evaluations are not used for performance-rating purposes or for the bestowal of rewards or incentives.

Nevertheless, there is no current policy on teacher evaluation *per se*. There is, however, a policy on school evaluation, from which certain inferences are being drawn with respect to teacher evaluations. The country's education authorities are attempting to establish a mechanism for conducting individual as well as school evaluations, which is one of the aims of the National School Evaluation System (SNED) and which allots incentive payments based on the quality of school performance as a whole. Their goal is to broaden the scope of the evaluation process: to take into account factors other than academic achievement, such as initiative or inventiveness, participation, ties to the surrounding environment, and the way in which each school is helping to promote equal opportunity. They are also seeking to develop incentive policies for schools and their staffs.

Under the SNED, the value of incentive pay is the same for all professional educators working in a given school, prorated according to the number of hours in their weekly work schedule under their respective employment contract. Thus, incentive payments reward good performance at the school-wide level. However, the law sets aside 10 percent of the incentive payment allotted to each school to enable its staff to independently decide whether to share it equally among all staff members or to bestow it on one or more teaching professionals regarded as particularly outstanding by the faculty as a whole. This provision establishes the concept of individual incentives.

An in-house Education Ministry report by Núñez (1996) outlines the goals sought by the joint Education Ministry/Teachers' Association committee in its study of a proposed amendment to the section that implements regulations for the Teachers' Act referring to performance evaluations of teaching professionals. These include making its language less punitive, promoting professional advancement, reducing the likelihood of arbitrary performance ratings, improving corresponding instruments, increasing the number of participants in evaluation processes, making good performance ratings the best form of protection for job security purposes, and allowing for repeated poor performance ratings to be used as a basis for making decisions in the best interests of educational quality.

There is no known written documentation on any experiences with individual evaluations of teacher performance, and this fact is noteworthy. We do know that certain educational institutions, particularly private schools, have had some rather disheartening experiences in this respect. These have only served to reinforce the reluctance of teaching professionals and their union to see any legislation enacted in this area.

El Salvador

The educational reform movement underway in this country has a great deal of momentum, in spite of the nation's complex socioeconomic situation produced by the war years and the postwar period and its efforts to consolidate the peace process.

Studies by Ottoniel (1990) on the status and profile of teachers trained during the 1980s and later studies by Reimers (1995) in the form of needs assessments underscore certain issues pertaining to teacher performance. These have prompted advocates of educational reform to recognize the need to establish teacher evaluation and performance-rating mechanisms.

The study by Ottoniel presents data on the human attributes and professional qualifications of teachers in three areas of the country (the west, the center, and the east) as two of many determining factors that figure in improving the quality of education. The data are used to construct a profile based on actual teachers and to pinpoint strategic problems that could be addressed by teacher evaluations.

An analysis of this study data reveals that, upon graduation, most new teachers are assigned to work at the lowest grade levels, which makes it difficult for them to develop

teaching experience. New teachers are aware of their serious shortcomings in regard to the writing and their unfamiliarity with step-by-step procedures for curriculum development and teaching methods. As a result they tend to improvise rather than plan their class work. Despite these problems, teachers are happy with their profession, except for the pay issue, because it allows them to get close to students and their families, as well as to their colleagues. However, the study reveals critical problems in the following areas:

- As far as teacher know-how, attitudes, and practices are concerned, there is a lack of coherence between current practices and existing regulations for the hiring of teachers;
- As far as classroom performance is concerned, there is too much reliance on improvisation and, rather than inventing creative learning situations attuned to the interests and needs of their students, teachers are merely parroting textbooks or syllabi and administering final exams that, at best, measure their students' ability to memorize facts and figures;
- Working conditions for teachers provide no motivation for upgrading their skills or filling in the gaps they face as professional educators.

These conditions have prompted experts in this area to make the following recommendations:

- Legislation and agreements with respect to the hiring of teachers need to be more strictly enforced;
- New teachers should be assigned to teach higher grades, lacking the knowledge and expertise needed to begin their teaching careers at lower grade levels; and
- Training, refresher training, and good performance ratings should be encouraged through appropriate incentives for promotions and higher pay.

Moreover, the study by Reimers lays the foundation for using a needs assessment to approach education as a national project. The study cites the need to invest in education and manpower training as a way to promote structural changes in the education system. It presents a series of suggestions mirroring the need to equip the education system with appropriate, efficient mechanisms for evaluating teacher performance.

The study emphasizes continuing to give top priority to basic education in order to strengthen the base of the education pyramid and, more specifically, to develop a teacher training system for primary school teachers with provisions for induction as well as inservice training. It recommends the creation and/or strengthening of what it refers to as "training workshops for teachers in service" and the promotion of forums enabling teachers to share their experiences. The study maintains that only through a process of trial and error will teachers succeed in making lasting improvements in their classroom techniques. What is this if not another form of performance evaluation?

Furthermore, the study underscores the fact that the nation has no specific mechanism for evaluating the performance of its education system as a whole. The current assump-

tion and expectation is that this function will be performed by the supervision department, which is overburdened with administrative responsibilities, and lacks the necessary technical expertise for this work. As a result, the Education Ministry is failing to provide leadership, nor is it monitoring their job performance.

While there are rules and regulations governing educational practice and the monitoring of teacher performance, they are scattered throughout different pieces of legislation. The legislation harbors a clearly protective stance toward the teaching profession, as reflected in an emphasis on job security and teachers' rights.

The promotion system puts a premium on seniority and preliminary professional training, while ignoring performance, which operates as a disincentive and helps perpetuate a state of inertia.

The General Education Act gives educators the right to promotion based on merit and qualifications, without actually defining what these terms mean. While promotions by seniority are clearly automatic, the General Regulations for Secondary Education broaden the definition of merit to include educational research and studies of educational problems, or the publication of textbooks. However, these merits are to be judged by an *ad hoc* committee, which is by no means a guarantee of impartiality. By the same token, the regulations fail to define what is meant by "the most efficient worker within the institution," referring to the filling of vacancies.

Considering the complexity of the situation described in these studies and taking into account the study data, the solution to the country's evaluation problems advanced in the National Commission on Education, Science, and Development (1995) proposal, which recommended the establishment of a reliable, systematic teacher performance evaluation system, is interesting, to say the least. The document contains the following outline of the distinctive features of the envisaged system, to help the Ministry move forward with its implementation:

- The system is designed to improve the efficiency and effectiveness of teachers, school principals, and assistant principals;
- The information furnished by the system is to be used for making decisions in regard to contract renewals, reassignments, and merit increases;
- The school principal is to be in charge of evaluating teacher performance;
- Evaluations are to be conducted at least once a year, with the participation of students and their parents;
- There is to be a Ministry presence, which will be responsible for conducting purposeful or random audits;
- The system should, ideally, be grounded in a new National Teacher Promotion Act establishing promotion policy and criteria, specifically in regard to teacher rating systems, bases for promotion (which should be by merit), and pay scales and salary differentials (which should be based on considerations of equity and fairness).

RECOMMENDATIONS

This study illustrates the importance of teacher performance evaluations in ongoing educational reform processes throughout the region. Educational quality hinges on a number of different factors, one of which is teacher performance. The importance of reforming and improving education systems and corresponding educational practices makes it imperative that education officials, teachers, and all stakeholders in the education process take immediate steps to update their practices and improve their performance in order to raise the quality of education.

However, given the heterogeneous nature of area countries and the different conditions and unique factors that need to be addressed in each specific locality, new approaches to ensuring the quality of these processes rather are needed rather than the development of standardized education systems.

Are such recommendations equally valid for all countries? Isn't it possible that actual conditions are even more disparate than we initially believed, which would call for different recommendations in line with the specific circumstances of each area? Isn't it better to suggest positions for the development of a policy package in the education sector than to simply make practical recommendations on what needs to be done? The last section of the paper presents precisely these sorts of recommendations.

Recommendations for research

1. We need to sharpen our knowledge of teacher performance evaluation practices and related issues, including their legal and regulatory framework, the administration of teacher evaluations at the school level, and union perceptions and reactions in this respect. Further studies of teacher performance evaluation methods, procedures, and strategies used in innovative programs and reform efforts would be especially valuable.

2. It is vital that we gain a better insight into the reasons behind union opposition to teacher evaluations. We must ascertain whether there is the same type of opposition to evaluations administered by parties within the same educational institution as there is to evaluations conducted by outside parties. We know there are standard practices and that a certain amount of teacher opposition to evaluation processes concerns the way in which the evaluation data are used, but there is no systematic research in this area.

3. There is very little research available in the area of the teacher evaluation process. No one knows for certain whether the information is nonexistent or if results are simply being withheld. In the latter case, it may well be that we are still clinging to a regulatory concept of evaluation, which can prop up highly vertical education systems but, by the same token, is totally inconsistent with the types of approaches that educational reform advocates are attempting to promote. This would be another interesting area for future research efforts.

4. It is important to promote the use of teacher evaluations for making decisions on salary-related matters, appointments, positions, and responsibilities, and to make these uses part of the same processes.

5. It would be equally helpful to share in the experiences of educational institutions that have developed their own approaches to teacher evaluation based on their own educational projects and on information furnished by experts as well as by research.

Policy recommendations

1. Education systems need to be equipped with a teacher evaluation system clarifying the objective for administering teacher evaluations and establishing performance evaluation criteria and related evaluation procedures. The notion of conducting evaluations for promotion purposes or for doling out rewards or punishments is unthinkable. All teacher evaluations, including evaluations of new teachers as well as teachers in service, should be geared to framing and strengthening strategies for improving educational quality.

2. Existing evaluation systems need to be flexible enough to allow the addition of new criteria inspired by educational practice. This will heighten teacher participation and give teachers a more active role in evaluating their own performance. Evaluation can help promote educational reforms only if those in charge of the everyday operation of educational institutions are involved in the study process. Thus, corresponding designs, proposals, and even procedures need to be based on the judgments and experiences of stakeholders regularly involved in the everyday realities of schools and educational processes.

3. As part of ongoing successful decentralization processes throughout the area, education ministries need to be actively involved in the creation of efficient mechanisms for the continuous technical support of each decentralized school system in its efforts to plan and design appropriate instruments and analyze resulting data.

4. Current mechanisms that may reflect obsolete concepts of education and evaluation need to be revamped.

5. It is vital that we continue to promote and to make information available on successful experiences with teacher performance evaluation processes in different countries.

NOTES

1. At a meeting of teachers specifically addressing this issue, one teacher pointed out "There's no such thing here. In practice, they're performed by the school principal and are subjective, according to his fondness for the subject. In other cases, they're more objective, but not conducted on a regular basis. In case of disciplinary problems with a particular teacher or a teacher

showing many students with failing grades, the principal will put pressure to bear, insisiting that "the teacher isn't working out."

2. This would later be embodied in the law, in the following provisions: "Performance evaluations and ratings of teaching personnel shall be recorded on forms to be designed by Headquarters subject to the approval of the General Administrator of Education based on the provisions of the performance evaluation and rating handbook, which shall also be drawn up by Headquarters" (the Teachers' Act, Chapter IX, Art. 75).

REFERENCES

Ahumada, P. (1992). *Evaluación de la Eficiencia Docente.* Chile: Universidad Católica de Valparaíso.

Arancibia, V. (1994). "Características de los Profesores Efectivos en Chile y su Impacto en el Rendimiento Escolar y Autoconcepto Académico." In *PSYKHE* 3, 2.

Arancibia, V. (1987). "Estado del Arte: Manejo Instruccional del Profesor en la Sala de Clases en América Latina." BRIDGES Project. Boston: Harvard University.

Avalos, B. and Haddad, W. (1981). *Reseña de la Investigación Sobre la Efectividad de los Maestros en el Africa, América Latina, las Filipinas, la India, Malasia, el Medio Oriente y Tailandia: Síntesis de Resultados.* Bogota, Colombia: CIID.

Cardemil, C. et al. (1991). "Factores que Inciden en el Mejoramiento de los Aprendizajes en Educación Básica." In *Cuadernos de Educación,* (CIDE. Santiago, Chile) 208.

Carranza, R. (1992). *Modelo de Evaluación para el Desempeño Docente.* University of Costa Rica: Education Department.

Connelly, F. M. (1989). "Evaluación de Profesores: una Revisión Crítica para una Práctica Reflexiva Supervisada." Paper presented at the Seminar on Teacher Training and Remuneration Policy. Montevideo, Uruguay.

Ezpeleta, J. (1989). *Escuelas y Maestros. Condiciones del Trabajo Docente en la Argentina.* Santiago, Chile: UNESCO-OREALC.

Fallas, A. et al. (1993). *Propuesta de Modelo de Evaluación del Desempeño Docente, con su Respectivo Manual, Normas, Procedimientos e Instrumentos: Caso del Sistema Educativo Saint Clare.* University of Costa Rica: Education Department.

Filp, J. (1994). "Todos los Niños Aprenden. Evaluaciones del P-900." In *Cooperación Internacional y Desarrollo de la Educación.* Chile: ASDI/AECI/CIDE.

Galo de Lara, C. (1990). *El Maestro como Orientador del Aprendizaje. Evaluación del Desempeño.* Guatemala:CINDEG.

García-Huidobro, J. E. (1994). "El Programa de las 900 Escuelas." In *Cooperación Internacional y Desarrollo de la Educación.* Chile: ASDI/AECI/CIDE.

Garro, G. (1988). "El Pago por Méritos y los Problemas de la Supervisión y la Evaluación de los Educadores." In *Revista de Educación de la Universidad de Costa Rica.* (San Jose, Costa Rica) 12, 1.

González Soler, A. (1980). "El Perfil del Profesor Eficaz como Base para la Evaluación de los Programas de Formación del Profesorado: Problemas y Perspectivas." In *La Investigación Pedagógica y la Formación de Profesores.* Madrid, Spain: C.S.I.C.

Mejía, M. (1987). *Movimiento Pedagógico: Una Búsqueda Plural de los Educadores Colombianos.* Bogota, Colombia: CINEP.

Namo de Mello, G. (1982). *Magisterio del Grau. Da Competencia Técnica al Compromiso Político.* Sao Paulo, Brazil: Autores Asociados.

Núñez, I. (1996). "Él Trabajo Docente: Condiciones y Regulaciones." Mimeographed document. Santiago, Chile: Ministry of Education.

Ordóñez, C. et al. (1996). *Propuesta de Organización del Sistema de Selección para Ingreso a la Carrera Docente.* Bogota, Colombia: ICFES.

Ottoniel, S. (1990). *Estudio del Perfil Real y de la Satisfacción Profesional del Maestro Salvadoreño Formado Durante la Década de 1980-1989.* El Salvador: ED-UCA.

Popham, J. (1980). *Problemas y Técnicas de la Evaluación Educativa.* Madrid, Spain: ANAYA.

Reimers, F. (1995). *La Educación en El Salvador de Cara al Siglo XXI. Desafíos y Oportunidades.* San Salvador, El Salvador:UCA Editores.

Rojas, C. and Castillo, Z. (1987). *Evaluación del Programa Escuela Nueva en Colombia.* Bogota, Colombia: Instituto SER de Investigaciones.

Rossenfeld, M. (1983). "Valoración de la Actuación de los Docentes: un Panorama." In *ACADEMIA* (Chile), 5-6. Academia Superior de Ciencias Pedagógicas de Santiago de Chile.

Rubilar, M. and Cuevas, S. (1996). "La supervisión pedagógica, una variable de apoyo al proceso de modernización de la educación. Resultados de una experiencia." In *Revista de Pedagogía* (Santiago, Chile), March 1996. FIDE.

Schön, D. (1987). *Educating the Reflective Practitioner.* London: Jossey-Bass.

Universidad Nacional de Colombia (1996). *Evaluación de Docentes.* Bogota, Colombia: UN/National Education Ministry Project.

Villa Sánchez, A. (1985). "La evaluación del profesor: perspectivas y resultados." In *Revista de Educación.* (Madrid, Spain), 277.

Section IV
EVALUATION OF THE ORGANIZATION OF EDUCATION

CHAPTER 9

EVALUATING THE PERFORMANCE OF INDIVIDUAL SCHOOLS

William J. Webster and Robert L. Mendro

The fourth part of this book is devoted to a topic that is becoming increasingly important, given the direction of new educational policies and the inexorable move toward greater participation by parents and communities in school government: school evaluation. The section begins with a chapter on school evaluation, which forms part of the tradition of social research and evaluation, and presents in detail the use of absolute and value-added components in the evaluation of schools, school districts, and state or national school systems. The authors illustrate their theoretical and methodological contributions by presenting the case of the Dallas, Texas, public schools in the United States, where value-added evaluation has been used to improve education and provide parents, teachers, and other members of the school community with knowledge about the progress of each component in the system.

INTRODUCTION

This monograph provides a framework for developing policy options to evaluate the performance of individual schools within a context of national educational reform, increased political decentralization, and local autonomy. First, a brief review of existing evaluation models is provided. This is followed by an exposition of a three-tier accountability system. The first tier focuses at the school level. Each school must have at its disposal data relative to the important educational outcomes to be measured and from which its standards are developed. The paper discusses the relationship between external standards and their measures and the school's own internal standards and measures and establishes the need for context, input, process, and product evaluation. The second tier is at the district or national level. The district or national agency sets the desired levels of accountability objectives and specifies the nature of individual school support and vehicles implementing this support. The third tier focuses on the development of unbiased measures of school effectiveness to be used in a value-added evaluation of schools. These measures take into account important school and student background

variables and prior student achievement to provide information on the degree to which a school is effective in enhancing the education of all of its students. The intent is to be able to design systems at school and national levels that examine both absolute and value-added components to provide comprehensive measures of school effects and effectiveness.

As schools move toward more autonomy and more control of their resources, the need for accountability becomes even more pressing. Site-based management carries with it a heavy site-based responsibility for assuring that students receive an adequate education. Accountability is the cornerstone on which a system of site-managed schools is built. The school-level accountability system must provide adequate data for site-level and oversight decision making as well as for accountability to the various clients of the school system. For purposes of this monograph, an educational evaluation study is one that is designed to assist some audience to judge and improve the worth of some educational object. It serves both decision making and accountability purposes.

ALTERNATIVE CONCEPTUALIZATIONS OF EVALUATION

Stufflebeam and Webster (1980) characterized thirteen different approaches to evaluation. The approaches were classified into categories of political orientation study types (labeled pseudo-evaluation), questions-orientation study types (labeled quasi-evaluation), and values-orientation study types (labeled true evaluation). Tables 1 through 3 present a concise overview of the three different study types and the thirteen different approaches to evaluation that operationalize them. These tables delineate the types of studies, the advance organizers for each study type, the purpose of each study type, the source of questions for each study type, the main questions posed by each study type, and the typical methods employed in each of the study types. For the questions-oriented and values-orientation study types, the pioneers are also delineated.

Most of the seminal work in educational evaluation was done in the 1960s and 1970s. Most recent work has centered on the development of appropriate methodologies to enhance evaluations rather than on the theory of evaluation *per se*. Stufflebeam, a prolific writer, has turned much of his attention to developing standards for educational evaluation rather than to embellishing his evaluation model. That model, the Context, Input, Process, Product model (CIPP), is clearly the most influential evaluation model in American schools (Webster, 1988). Two works, *Evaluation Models* (Madaus, Scriven, and Stufflebeam, 1983) and *Meta-Evaluation of School Evaluation Models* (Gallegos, 1994), provide presentations and reviews of various evaluation models. The first publication provides original essays by many of the major evaluation theoreticians. In the second publication, the author presents fifty-one different evaluation models. Stufflebeam (1996) provides an excellent framework for the evaluation of students, programs, and personnel.

It is our view that the most appropriate approaches to evaluation involve values-oriented study types and that the most useful values-oriented study types are decision-oriented studies. We hold this view because decision-oriented study types focus not only on accountability, but also on providing useful information for decision making and improvement. Accountability without information for improvement is not very useful if one's purpose is to use accountability to improve the system being evaluated. It is also our view that the design of any evaluation system should be grounded in program evaluation standards in relation to utility, feasibility, propriety, and accuracy (Joint Committee on Standards for Educational Evaluation, 1994). If these standards are not met, stakeholders in the schools may be forced to make decisions based on inaccurate, invalid, incomplete, or incomprehensible information.

Table 1. Political-Orientation Study Types (Pseudo-Evaluation)

Approach	Political Orientation (Pseudo-evaluation)	
Definition	Studies that promote a positive or negative view of an object irrespective of its worth.	
Study Type	Politically Controlled Studies	Public relations inspired studies
Advance Organizer	Implicit or Explicit Threats	Propagandist's information needs
Purpose	To acquire, maintain, or increase a sphere of influence, power or money	To create a positive public image for an object
Source of Questions	Special interest groups	Public relations specialists and administrators
Main Questions	What information would be best to report or withhold in a projected confrontation?	What information would be most helpful in securing public support?
Typical Method	Covert investigations and simulation studies	Biased use of surveys, experiments, and "expert" consultants

Table 2. Questions-Orientation Study Types (Quasi-Evaluation)

Approach	Questions Orientation (Quasi-Evaluation)				
Definition	Studies that address specified questions whose answers may or may not assess an object's worth				
Type of study	Objectives-based Studies	Accountability Studies	Experimental Research Studies	Testing Programs	Management Information Systems
Advance Organizers	Objectives	Personnel/ institutional responsibilities	Problem statements, hypotheses, and questions	Areas of the curriculum, published tests, and specified norm groups	Program objectives, activities, and events
Purpose	To relate outcomes to objectives	To provide constituents with an accurate accounting of results	To determine the causal relationship between specified independent and dependent variables	To compare the test performance of individual students and groups of students to select norms	To continuously supply the information needed to fund, direct, and control programs
Source of Questions	Program developers and managers	Constituents	Researchers and developers	Test publishers and test selection committees	Management personnel
Main Questions	Which students achieved which objectives?	Are those persons and organizations charged with responsibility achieving all they should achieve?	What are the effects of a given intervention on specified outcome variables?	Is the test performance of individual students at or above the average performance of the norm group?	Are program activities being implemented on schedule, at a reasonable cost, and with expected results?
Typical Methods	Analysis of performance data relative to specified objectives	Auditing procedures and mandated testing programs	Experimental and quasi-experimental designs	Selecting, administering, scoring, and reporting standardized tests	System analysis PERT, CPM, PPBS, computer-based information systems, and cost analysis
Pioneers	Tyler (1949)	Lessinger (1970)	Campbell and Standley (1963)	Lindquist (1951)	Cook (1966)

Table 3. Values-Orientation Study Types (True Evaluation)

Approach	Values Orientation (True Evaluation)					
Definition	Studies that are designed primarily to assess some object's worth					
Study Type	Accreditation /certification studies	Policy studies	Decision-oriented Studies	Consumer-oriented Studies	Client-centered Studies	Connoisseur-based Studies
Advance Organizer	Accreditation /certification guidelines	Policy issues	Decision situations	Societal values and needs	Localized concerns and issues	Evaluators' expertise and sensitivities
Purpose	To determine whether institutions, programs, and personnel should be approved to perform specified functions	To identify and assess the potential cost and benefits of competing policies for a given institution or society	To provide a knowledge and value base for making and defending decisions	To judge the relative merits of alternative educational goods and services	To foster understanding of activities and ways they are valued in a given setting and from a variety of perspectives	To describe, appraise, and illuminate an object critically
Source of Questions	Accrediting/ certifying agencies	Legislators, policy boards, and special interest groups	Decision makers (administrators, parents, students, teachers), their constituents, and evaluators	Society at large, consumers, and the evaluator	Community and practitioner groups in local environments and educational experts	Critics and authorities
Main Question	Are institutions, programs, and personnel meeting minimum standards; and how can they be improved?	Which of two or more competing policies will maximize the achievement of valued outcomes at a reasonable cost?	How should a given enterprise be planned, executed, and recycled in order to foster human growth and development at a reasonable cost?	Which of several alternative consumable objects is the best buy, given their costs, the needs of the consumers, and the values of society at large?	What is the history and status of a program and how is it judged by those who are involved with it and those who have expertise in program areas?	What merits and demerits distinguish an object from others of the same general kind?
Typical Methods	Self-studies and visits by expert panels to assess performance in relation to specified guidelines	Delphi, experimental and quasi-experimental design, scenarios, forecasting, and judicial proceedings	Surveys, needs assessments, case studies, advocate teams, observation, and quasi-experimental and experimental design	Checklists, needs assessments, goal-free evaluation, experimental and quasi-experimental designs, modus operandi analysis, and cost analysis	Case study, adversary reports, sociodrama, responsive evaluation	Systematic use of refined perceptual sensitivities and various ways of conveying meaning and feelings
Pioneers	College Entrance Examination Board (1901)	Coleman (1966)	Cronbach (1963), Stufflebeam (1966, 1967)	Scriven (1967)	Stake (1967)	Eisner (1976)

CONSIDERATIONS IN DEVELOPING A SCHOOL EVALUATION MODEL

In developing an evaluation model for schools that meets the evaluation standards alluded to above, the authors have relied heavily on the work of Stufflebeam et al. (1971), Scriven (1967), Stake (1967), and Provus (1971) in conjunction with advances in value-added methodology, specifically hierarchical linear modeling, as described by Bryk, et al. (1988a), Bryk and Raudenbush (1992), Bock (1989), and Goldstein (1987). A brief description of each of these components follow. The astute reader will note a definite orientation toward decision-oriented studies.

Evaluation Models

Probably the most comprehensive evaluation model ever developed was Stufflebeam's CIPP model (Stufflebeam et al., 1971). Evaluation was defined as the process of delineating, obtaining, and providing useful information for judging decision alternatives. The model identified four major types of evaluation: context evaluation to feed planning decisions, input evaluation to feed programming decisions, process evaluation to feed implementing decisions, and product evaluation to feed recycling decisions.

Context evaluation provides a rationale for determining educational objectives by defining the relevant environment, describing desired and actual conditions of the environment, identifying unmet needs, and diagnosing problems that prevent needs from being met. Input evaluation assesses relevant capabilities of responsible agencies and identifies strategies for achieving the objectives determined through context evaluation as well as suggesting designs for implementing selected strategies. Once a strategy has been selected, process evaluation provides periodic feedback to persons concerned with the implementation of plans and procedures to predict or detect faults in procedural design or implementation so that interim adjustments may be made if warranted. Finally, product evaluation provides interim and final assessment of the effects of educational programs. That is, product evaluation assesses the effects of the strategies selected through input evaluation to meet the needs identified by context evaluation. Such assessment is completed in light of process evaluation data.

Scriven (1967) conceptualized an extremely straightforward and widely accepted evaluation framework. It is not nearly as comprehensive as the CIPP Model and is largely concerned with the process-product portion of Stufflebeam's structure. According to Scriven, the major goal of evaluation is to make credible judgments relative to the merit and worth of educational programs. Within a discussion of methods of accomplishing this goal, he introduced the concepts of formative and summative evaluation.

The focus of formative evaluation is upon program or school improvement. Thus, formative evaluation attempts to provide feedback to program personnel with the goal of upgrading or improving an educational program while it is in the developmental stage. In the CIPP vernacular, interim product and process data provide formative evaluation information to program personnel.

The focus of summative evaluation is upon the determination of the ultimate worth of a program or project. This type of evaluation should be implemented at that stage in a program's life where it has reached some stability. Summative data feed recycling decisions: that is, as a result of summative evaluation information, a program may be terminated, restructured, continued, or expanded. In the CIPP vernacular, final product evaluation information, interpreted in consideration of context, input, and process data, is used to draw summative conclusions about the merit and worth of an educational program and feed recycling decisions.

Stake (1967) suggested that evaluation ought to be concerned with three classes of conditions: antecedents, transactions, and outcomes. Antecedents are defined as those conditions that exist prior to program implementation, i.e., the educational context. Transactions are interactions between students, teachers, and materials. Outcomes are defined as the intended products of transactions.

Stake further suggested three classes of activities. The first involves providing assistance to program staff in generating a clear statement of the program or project rationale. The second involves the generation of descriptive data. Descriptive data include statements regarding intended and actual antecedents, transactions, and outcomes. Thus a check of the congruence between planned and observed antecedents, transactions, and outcomes can be made. Stake also suggested an examination of the contingencies within intended (logical contingency analysis) and observed antecedents, transactions, and outcomes. The contingency analysis within intended conditions is similar to CIPP's input evaluation, while that within observed data attempts to identify cause and effect relationships between antecedents, transactions, and outcomes.

The third class of activities involves the generation of judgments about the worth of educational programs. Stake suggested that such judgments be made on the basis of both absolute and relative criteria and by a variety of individuals. In other words, programs should be assessed both in terms of the degree to which they attain absolute, and sometimes arbitrary, goals and of the degree to which they attain those goals relative to other programs with similar goals or objectives.
Provus (1971) suggested that all projects move through design, installation, process, and product stages. During each stage the evaluator must delineate, in conjunction with project staff, a set of standards that can be used as a basis for comparison with program performance. It is the evaluator's function to make comparisons between standards and performance, to identify discrepancies at each stage, and to report those discrepancies to project management, which has the option of terminating the program, proceeding to the next stage, or modifying the program in some way. The product of the design stage is a set of standards used to judge the effects of program efforts in each of the three succeeding stages. At every stage the object of the evaluation is to provide useful data for decisions about program improvement or recycling.

While the principal focus of the four evaluation approaches is program evaluation, the translation to school evaluation is straightforward. This translation is made later in this monograph.

The utility of unadjusted versus value-added outcomes in school evaluation

To this point, various systems of school evaluation have been discussed that show the different models available and the ways in which context, input, and process are delineated and are related to outcomes. Outcomes have been considered at only the unadjusted level. Some educational outcomes of value are, and should be, unadjusted. However, the use of unadjusted outcomes without awareness of their characteristics can result in biased and misleading school evaluation.

The purpose of this section is to define and discuss the use of value-added outcomes in school evaluation in place of unadjusted outcomes. This section will show the types of criteria against which all outcomes, unadjusted and value-added, must be measured and will provide some examples of misinterpretation of school effect through the use of unadjusted outcomes.

For explication, let us begin with examples of two unadjusted outcomes and the goals based on them. Assume in the first example that a school has a dropout rate of 15 percent and a goal is set for the school to reduce it to 13 percent. Assume in the second example that every school in a system has at least 50 percent of its students reading at grade-level and a system goal is set for every school to have 60 percent of its students reading at grade-level. On the surface these seem like realistic uses of unadjusted outcomes and seem to make rational goals.

Now consider both examples further. Assume in the first example that the population of students on which the school draws has a general history of a dropout rate ranging between 9 percent and 11 percent. Now the unadjusted outcome of 15 percent becomes undesirable and a goal of 13 percent too high a rate. A more appropriate goal for the school might be to reduce the rate to 12 percent in the first year and below 11 percent thereafter. In the second example, assume that in the past two years every school has had at some recent point 60 percent of its students reading at grade level. An unadjusted goal for every school might be to have at least 65 percent of its students reading at grade level every year.

The observant reader will note the caveats attached to each of these examples. The goals and outcomes are conditioned on the past performance of the underlying populations of students. Reconsider the two examples with different underlying conditions and these same unadjusted outcomes and their accompanying goals can rapidly become inappropriate for different reasons. In the first example, the school has held its dropout rate to 15 percent with a given population of students. If we now assume that the dropout rate for the population of students on which the school draws is 18 percent to 20 percent and that no other school has been able to reduce the dropout rate for a similar population to below 17 percent, the goal of 13 percent for that school may well be unrealistic for a different reason, i.e., the school's rate of 15 percent may be an example of current best practice. In all likelihood, the unadjusted rate of 15 percent is an excellent outcome. The school is to be commended for keeping it at that rate and should be sharing its techniques with the remaining schools. The unadjusted rate of 15

percent will remain an excellent outcome until it can be demonstrated that more effective techniques can reduce the underlying population rate below 15 percent.

In the second example, the system goal was 60 percent of students reading at grade level for every school. But now assume that this year is the first year ever that all schools have reached the goal of 50 percent of students reading at grade level. Assume also that several schools have never dropped below 75 percent of their students reading at grade level. Now the system goal of 60 percent may be more appropriately 55 percent for some schools and 80 percent for other schools.

Clearly in both of these examples, context matters. Outcomes and goals must be considered in light of this context. As Glass (1978) argues, the outcomes and the goals or standards developed from them are relative to the existing performance of their specific groups of students. This example also illustrates one of the major problems associated with unadjusted outcomes. Absolute goals, based on these outcomes, are established without any thought as to whether there is any probability of making the goals. Webster and Mendro (1995) discuss this problem at length and show that achievable goals can be set based on unadjusted outcomes. The most pressing problem is that the public or higher administration would like to see massive progress and typically feels that carefully constructed, incremental goals present a problem of "low expectations." Extreme, unrealistic goals, with little probability of attainment, are generally more satisfying to those outside of the school. Another cogent problem that we have noted is that few educators are able to make a direct translation of any goals based on academic outcome measures into a plan of action for school improvement. This is a pervasive problem that affects all educational improvement efforts and is discussed further in a later portion of this monograph.

Regardless of how unadjusted data are used in goal setting, the question of evaluating educational progress in a fair and precise manner remains. How are we to determine the appropriate contexts for evaluating educational outcomes at the school level? How can we determine whether an outcome is actually improvement or is only the result of typical progress for a defined population?

The answers to these questions and many like questions lie in the rapidly developing field of value-added assessment of educational outcomes. With value-added systems, conditions outside the control of a school are held constant for all schools in a group. The effects obtained by each school are measured on a common metric and the results compared. In essence, all schools in the group are set at a common baseline and the critical element then becomes whether, relative to other schools, a school has had a positive or a negative effect on its students. Has it added value to or subtracted value from the base? This common baseline helps answer the question of sorting out improvement from typical performance.

These systems can be constructed with a variety of similar methodologies, all of which provide preferable alternatives to systems based on unadjusted outcomes. Although we will discuss a system that has been carefully researched and fine-tuned to eliminate

many small biases, most of the regression-based value-added systems are far better than the alternative unadjusted systems. The essence of all of these systems is to eliminate known factors that affect school outcomes but are not possible for the school to control. At that point, improvement can be identified. Further, when these systems are designed properly, they offer the best chance of adjusting for effects of variables at the student level that are not explicitly included in the known factors. In other words, they help control to some extent all factors that are not under the control of the schools.

Some of the known factors that affect school progress but are outside of the control of the school include, but are not limited to, socioeconomic status, gender, language proficiency, ethnicity, and the existing ability levels of entering students. The criteria by which all systems for evaluating schools need to be judged are the degrees of relationship between these factors and the resulting measures of effectiveness. Using these relationships as the criterion measures for the models is rare based on our extensive search of the literature on school effectiveness systems.

Systems that employ unadjusted outcomes or testing programs as their basis are too highly correlated with the existing factors just delineated. As noted in Jaeger (1992) and Webster et al. (1995), these types of systems are biased against schools with larger proportions of minority and low socioeconomic status students and are biased in favor of schools that contain larger proportions of white and higher socioeconomic status students. A comparison of Dallas schools ranked with a value-added system and schools ranked under an unadjusted accreditation system showed clearly that effective schools that also did well in the unadjusted system had higher proportions of white students and affluent families. It also demonstrated that schools that were very effective with their populations, but had high proportions of minorities and economically disadvantaged students, performed at lesser levels on the unadjusted system. Finally, the same study showed that correlations of school effectiveness rankings from an unadjusted system with these demographic factors is unacceptably high, ranging as high as, in the absolute value, .90 at the student level and .65 at the school level (Webster et al., 1995).

The essence of these arguments is that with unadjusted outcomes, schools are ranked primarily on the types of students who enter the schools, rather than on the education that the schools provide. Use of unadjusted outcomes in the comparison of schools and programs confounds the differences in populations of students and how they are selected into their schools and programs with the difference the schools and programs make. Schools and programs that draw on higher-scoring students receive the benefits of this bias before their students start their first lesson. Schools and programs that must deal with lower-scoring students must overcome this bias before they can begin to show an effect.

Required elements of a school evaluation model

In implementing a school evaluation system, a number of required elements must be met if the system is to provide useful estimates of school effect. These elements are presented and discussed in more detail below. To summarize, a school evaluation model must:

1. be value-added.
2. include provisions for context, input, process, and product evaluation.
3. be under the control of a representative group of school constituents.
4. include a broad array of outcome measures.
5. be based on the students continuously enrolled in a school.
6. include a provision to test virtually all eligible students.
7. include prior measures of all outcome variables at both the student and school level (fairness).
8. control for student and school level contextual variables over which the schools have no control (fairness).
9. provide information for improvement.

First, the *value-added requirement*, as previously discussed, means that schools must only be held accountable for the progress or lack of progress that they make with the students assigned to them. Thus, some measure of student gain or improvement is required. This requirement mitigates against the exclusive use of objectives-based studies, testing programs, and management information systems in evaluating schools.

It may seem simple, but the concept of value-added implies value added to *something*. In order to determine what a school adds to a student's education, there must be an initial measure related to the outcome measure. A system that attempts to remove the effects of related measures without removing the effects of student ability will result in high correlations between initial student ability measures and the effectiveness measures (Webster et al., 1995). Inclusion of measures such as socioeconomic status in and of themselves, without a prior measurement of achievement, does not control the effects of prior student achievement to any great degree. The prior measure does not have to be, indeed cannot always be, the same measure as the outcome measure. The only requirement is that it be correlated with the outcome measure and be related to the measure through directly similar skills or through underlying skills.

In the United States, state departments of education have taken a leadership role in attempting to implement school evaluation and accountability systems. Forty-six states have accountability systems that feature some type of assessment. Twenty-seven of these systems feature reports at the school, district, and state level; three feature school-level reports only; six feature reports at both the school and district level; seven feature reports at the district and state level; two feature reports at the state level only; and one is currently under development (Council of Chief State School Officers, 1995). When one reviews these systems, it is obvious that their designers are not familiar with the literature on value-added systems since only two states, South Carolina (May, 1990) and Tennessee (Sanders and Horn, 1995), have used appropriate value-added statistical methodology in implementing such systems. Most of the rest tend to evaluate students, not schools or districts, and generally cause more harm than good with systematic misinformation about the contributions of schools and districts to student academic accomplishments. In attempting to eliminate bias, a number of states have gone to non-statistical grouping techniques, an approach that has serious limitations when there is consistent one-directional variance on the grouping characteristics within groups.

Fennessey and Salganik (1983) proposed a model for analyzing instructional program effectiveness within the context of gain scores. The rescaled and adjusted gain score (RAGS) index equalized aggregate net bias from responsiveness to instruction, regression-to-the-mean, and boundary artifacts in all program groups. A crucial assumption to this approach is that any group of students with similar pretest scores will have similar rates of learning and will be subject to the same degree of regress-to-the-mean. While the RAGS procedure is appropriate for program evaluation, it would be difficult to apply in a situation where one is attempting to determine the relative effectiveness of schools with very different student populations.

Another approach to the estimation of added value, which has received generally widespread acceptance among educational researchers, involves the aggregation of residuals from student-level regression models (Aiken and West, 1991; Bano, 1985; Felter and Carlson, 1985; Kirst, 1986; Klitgard and Hall, 1973; McKenzie, 1983; Millman, 1981; Saka, 1989; Webster and Olson, 1988; Webster, Mendro, and Almaguer, 1994). These techniques can incorporate a large number of input, process, and outcome variables into an equation and determine the average deviation from the predicted student outcome values for each school. Schools are then ranked on the average deviation. Some advantages of multiple regression analysis over other statistical techniques for this application include its relative simplicity of application and interpretation, its robustness, and the fact that general methods of structuring complex regression equations to include combinations of categorical and continuous variables and their interactions are relatively straightforward (Aiken and West, 1991; Cohen, 1968; Cohen and Cohen, 1983; Darlington, 1990).

Finally, hierarchical linear modeling (HLM) provides estimates of linear equations that explain outcomes for group members as a function of the characteristics of the group as well as the characteristics of the members. Because HLM involves the prediction of outcomes of members who are nested within groups, which in turn may be nested in larger groups, the technique is well suited for use in education. The nested structure of students within classrooms and classrooms within schools produces a different variance at each level for factors measured at that level. Bryk et al. (1988b) cited four advantages of HLM over regular linear models. First, it can explain student achievement and growth as a function of school-level or classroom-level characteristics while taking into account the variance of student outcomes within schools. Second, it can model the effects of student characteristics, such as gender, race-ethnicity, or socioeconomic status, on achievement within schools or classrooms and then explain differences in these effects between schools or classrooms using school or classroom characteristics. Third, it can model the between and within-school variance at the same time and thus produce more accurate estimates of student outcomes. Finally, it can produce better estimates of the predictors of student outcomes within schools and classrooms, by "borrowing" information about these relationships from other schools and classrooms. HLM models are discussed in the literature under a number of different names by different authors from a number of disciplines (Bryk and Raudenbush, 1992; Dempster, Rubin, and Tsutakawa, 1981; Elston and Grizzle, 1962; Goldstein, 1987; Henderson, 1984; Laird and Ware, 1982; Longford, 1987; Mason, Wong, and Entwistle, 1984; Rosenberg, 1973).

The third important required element of a school evaluation model is that the characteristics of the model should be under the control of a representative group of school constituents or stakeholders. These stakeholders should constitute a governing body that makes all final decisions regarding the system. This group might include parents, community members, business representatives, school administrators, teachers, students, board members and, in systems under government aegis, members or representatives appointed by the governing agency (at the city, state, or national level). Because the simplest systems involve a degree of mathematical complexity, statisticians and data analysis specialists will most likely have to advise the body. It is especially helpful if one or more statisticians are among the community members or business representatives, since the other members can receive an independent confirmation of the mathematics from one of their own members. That independent confirmation is extremely useful in building trust in the system among the group.

In Dallas, the governing body was designated the Accountability Task Force (Dallas Public Schools, 1996; Webster et al., 1997a). It is composed of parents, community members and business representatives, principals and teachers, and administrators representing the Superintendent. The Department of Research and Evaluation acts in an advisory capacity to provide statistical analysis services to the task force.
The Accountability Task Force makes the following decisions about the Dallas value-added assessment system:

- Selecting of outcome, dependent, and concomitant variables in the system
- Developing and overseeing the rules and procedures of the system
- Hearing all appeals of procedures and results by schools
- Advising the General Superintendent and the Board of Education on all external decisions regarding the system.

Since the Dallas system involves monetary rewards for the most effective schools and the Board of Education uses the results of the system to determine low-performing schools, the appeals and advisory functions of the Accountability Task Force are extremely important. No school or its leaders want their fate in such a system to be decided without an opportunity for some recourse if they feel the need for such.

A fourth required element, multiple outcome variables, is essential to any school evaluation or accountability system. It is not appropriate to report only the results of norm-referenced or criterion-referenced tests and call that an accountability system. Schools are responsible for many student outcomes. In addition to norm-referenced and criterion-referenced test results, important outcomes of schooling include, but are not limited to, student writing samples and performance measures, promotion rates, student attendance rates, graduation rates, dropout rates, enrollment in prehonors and honors courses and advanced diploma plans, and student success on advanced placement exams and in college or their chosen profession.

The larger the number of variables, the less a school can concentrate on a single test and the more it must concentrate on a general broad education with only limited time spent

teaching to any given test measure. We firmly believe that students must be taught some test-taking skills, but that the majority of the school curriculum should go above and beyond a particular test or tests.

In a study of the Dallas Mathematics program currently in preparation, curriculum materials by private publishers designed to assist schools in passing the state test were found in many schools. One set instructed fourth-grade teachers not to teach a subject found in all of the state-approved materials because it is not included on the state test. It argued that teachers should only teach what will be on the test in order not to confuse students (Bearden, 1997). This sort of travesty is what should be avoided through a broad array of variables in a value-added system.

After variable selection, the next most important component is that of continuous enrollment. This deals with the students and outcomes that will be counted in the system. In school evaluation, a basic component noted earlier in the discussion of the CIPP model was process evaluation. It is inappropriate to attribute the results of a program to a treatment if the treatment was not implemented. Similarly, in school effectiveness systems, if a student moves to a school long after the start of a school year, the effects associated with that student should not be attributed to that school. In other words, the effectiveness of schools should be determined on the performance of the students who were enrolled in that school for the majority of the school year. To base results on students who transfer into the school the week before testing can only be misleading. It is therefore necessary to select a minimum period during which a student must be enrolled in a school before the results for that student will be attributed to that school.

Fifth, school evaluation systems must be based on cohorts of students. Scores of different students over time are subject to fluctuations that have nothing to do with school effect. In addition, the system must be designed so that schools derive no particular advantage from starting with high-scoring or low-scoring students. This requirement demands the use of statistical methodology to predict and interpret student outcomes.

The sixth necessary component of a value-added system is a percent-tested rule. A school can hide students from testing when there is no pressure to avoid doing so. For tests with makeup periods, a rule requiring a specific percentage of eligible students to be tested should be in place. We recommend 95 percent of eligible students tested based on our experiences over a decade of required testing in Dallas (Dallas Public Schools, 1996). For tests with no makeup periods, such as the Texas TAAS test, schools should be expected to meet their average daily attendance minus a predetermined percentage. The authors recommend ADA minus 2 percent. These rules provide an assurance that schools are not attempting to subvert the system by selective testing. The authors note, however, that schools typically attempt to circumvent the testing rule by withholding low-scoring students. In a value-added system, this strategy can backfire because a low-scoring student can add as much to an effectiveness measure as a high-scoring student. It is only when an unadjusted system is used or a school systematically underserves low-

scoring students that the strategy of withholding low-scoring students has a payoff. A percent-tested rule will control the former situation and, in the latter situation, the value-added assessment information can be explicitly analyzed by student group to determine whether a group is being underserved.

Also essential is that the evaluation system be fair. This means that the influence of important student background variables over which the schools have no control must be controlled for. Such variables include student ethnicity, gender, primary language proficiency, socioeconomic status, and any other contextual variables that can be demonstrated to be related to the outcomes of interest. School-level variables that influence achievement but are not under the control of the schools must also be considered. School-level fairness variables might include student mobility, overcrowding conditions, average family income and education level, percentage of low socioeconomic students, percentage of various ethnicities, and percentage instructional days lost to medical disability leaves and unfilled teacher vacancies.

THE THREE-TIERED SYSTEM OF THE SCHOOL EVALUATION MODEL

The school evaluation model that is proposed in this monograph, and has been successfully implemented elsewhere, is a three-tiered system. The first tier focuses at the school level and is designed to hold each school accountable for most aspects of its operation. Greater school autonomy provides the promise of maximization of resources but carries with it the possibility of increased harm to individual constituencies. Because of this possibility, accountability must be foremost in plans for increased site-based decision making. Each school must be provided with useful data for decision making but must also be accountable for decisions. School Improvement Plans (SIPs) are the vehicles through which schools focus their efforts on improvement and provide the necessary information for evaluation and accountability.

The second tier focuses at the district level and is implemented through a District Improvement Plan (DIP). The DIP establishes the desired objectives of instruction and desired performance levels, and specifies how central office divisions support the schools. Depending on the manner in which a country's education system is organized, the DIP could be a regional, state, or national improvement plan.

The third tier of the system involves school improvement or effectiveness indices. These indices take into consideration important student background variables and provide information on how well schools function with the students that they serve. The SIP and DIP components of the system focus on the end products of schooling while the school effectiveness indices (SEIs) provide a value-added component to the system.

It cannot be emphasized enough that the two most important characteristics of a school evaluation or accountability system are fairness and usability. Educators who espouse the accountability movement have a right to know that the standards by which they are judged are fair and objective. It is essential that these systems also provide useful data

for decision making and improvement. The system outlined in this monograph incorporates fairness as defined by the Program Evaluation Standards (Joint Committee on Standards for Educational Evaluation, 1994) and the Standards for Educational and Psychological Testing (AERA, APA, NCME, 1985). It also provides data for decision making and improvement at each level of the system.

In discussing the accountability system, it is important to note that it is designed to be implemented in the context of site-based management. The Dallas Public Schools are implementing school-centered education through the Yale Child Study Center School Development Program (Comer, 1988). Under the model, the principal, parents, and staff are involved in school decision making and governance through a school-community council (SCC) that makes all relevant decisions about school operations. A number of committees can exist at each school but the SCC and its committees must take responsibility for curriculum, instruction, assessment (other than systemwide accountability measures), parental and staff skills development, school-community socialization and interaction, public relations, evaluation, and modification. At the high school level, these committees include students.

Regardless of the structure, the evaluation functions that are undertaken at the school level include the development of a SIP; the interpretation of formative data for use in problem-solving and of summative data for use in refocusing priorities, programs, and resources; the development of an implementation record of the various projects and programs within the school, including monitoring the implementation of the SIP; and the coordination of all school-based action research. Central office research staff, be they at the district, regional, state, or national level, must provide school personnel with training regarding how to accomplish many of the aforementioned tasks.

In other words, the District's School-Centered Education Plan focuses control of most available resources and all instructional decisions at the local school level (Edwards, 1991). The only decisions that school-level committees are not empowered to make are those involving the nature and magnitude of outcomes for which they are being held accountable. An extremely important step in the school improvement process is the determination of performance indicators that will inform educators, parents, and community members whether or not students are making satisfactory progress in the key developmental pathways that they believe are critical for academic learning. These performance indicators are determined by an Accountability Task Force and influenced by the state's Academic Excellence Indicator System. The Academic Excellence Indicator System is the basis for school accreditation in Texas. The accountability indicators are consistent across the three tiers of the accountability system.

The key to the success of the system described in this paper is the Accountability Task Force, a twenty-seven-member committee appointed by the Board of Education, and charged with the responsibility of overseeing the District's accountability system. The membership includes four elementary teachers, three middle school teachers, four high school teachers, four principals, four parents, five members of the business community, and three central office administrators. In addition, each of the various employee

organizations has an ex officio member on the task force. The task force deals with many aspects of the accountability system including methodology, testing, performance variables, and the rules for financial awards related to the accountability system. The Accountability Task Force also hears concerns or grievances. The formation of a group of stakeholders to oversee the accountability system is fundamental to the operation of a fair and equitable system.

First tier: the school improvement process

The major instrument of school improvement must be the School Improvement Plan (SIP). SIPs are organized around outcome targets that focus directly on the school's priorities. These priorities must relate directly to the priorities of the district, region, or state. Each school must do its part in meeting the important objectives of the district or state. SIP targets might include 1) student performance in language arts (vocabulary, reading, oral competency, and writing skills); 2) student performance in mathematics (problem-solving, concepts, and computational skills); 3) student performance in social studies; 4) student performance in science; 5) parental and community involvement in the schools; 6) student promotion and course passing rates; 7) student enrollment in advanced courses, diploma plans, and honors programs; 8) student graduation rates (dropout rates); 9) student college entrance test participation and performance; 10) student attendance; 11) teacher attendance; and 12) school climate and safety.

As part of its SIP, each school should develop strategic plans of action for each target. Each plan of action should include the following elements:

1. *Need:* a needs-assessment summary describing the current status of the target.
2. *Goal:* reference to the school's minimum accountability objectives or other standard of performance that will be met by implementing the plan. These are directly related to district or state goals.
3. *Narrative of Strategy:* a summary of what will be done to address the target.
4. *Waiver:* a specification of waivers from district or state policy required to implement the strategy.
5. *Activities/Timelines/Personnel Responsible:* activities, corresponding timelines, and personnel responsible for meeting the school's targets.
6. *Monitoring:* the methodology for directing, assessing, adjusting, and documenting formative activities to meet the goal.
7. *Resource Implications:* a summary of the distribution (e.g., monies, personnel) changes required to implement the strategies.

Figure 1 shows an example of SIP targets. Each school receives its own data on each of these targets and is responsible for achieving its targeted outcomes. The targets are criterion-referenced in the sense that schools have absolute goals and concentrate available resources on attempting to achieve those goals. While the data in Figure 1 are included in the SIP, there is a vast array of backup data given to schools that provide the necessary detail to diagnose student weaknesses. (Figure 5 provides examples of these data.)

Figure 1. Example High School Profile Featuring SIP Targets

Outcome Variables	Baseline	Year 1	Year 2	Year 3	Year 4
1. LANGUAGE ARTS					
TAP Reading (NCE)	46	48	50	52	54
TAAS Reading (Percent Passing)	55	58	61	64	67
TAAS Writing (Percent Passing)	80	80	80	80	80
ACP's (Percent Correct)					
English I	69	72	75	78	81
English II	58	62	66	70	74
English III	64	66	68	70	72
English IV	80	81	82	83	84
2. MATHEMATICS					
TAP Mathematics (NCE)	50	51	52	53	54
TAAS Mathematics (Percent Passing)	48	49	50	51	52
ACP's (Percent Correct)					
Algebra I	49	53	57	60	63
Algebra II	55	57	59	61	63
Geometry	70	70	70	70	70
Trigonometry	72	73	74	75	76
Pre-Calculus	75	75	75	75	75
Calculus	85	86	87	88	89
3. SOCIAL STUDIES					
ACP's					
World History	60	62	65	67	70
World Geography	68	70	72	74	76
U.S. History	70	73	76	79	82
Economics	80	81	82	83	84
4. SCIENCE					
ACP's					
Biology	65	67	69	71	73
Physics	68	70	69	71	73
Chemistry	75	76	77	78	79
5. PARTICIPATION OF PARENTS/COMMUNITY					
Volunteer Hours/Student	20	25	30	35	40
Percent Parent Involvement	60	65	70	75	80

6. STUDENT PROMOTION AND COURSE PASSING RATES					
Course Passing Rate	55	14	16	18	20
7. STUDENT ENROLLMENT IN ADVANCED COURSES, etc.					
Percent Enrolled in Advanced Courses	12	14	16	18	20
Percent in Advanced Diploma Plans	6	10	14	18	20
Percent in Honors Programs	15	17	20	22	25
8. STUDENT GRADUATION RATES					
Graduation Rate	50	53	56	59	62
Dropout Rate	8	7	6	5	4
9. COLLEGE ENTRANCE EXAMS					
Percent Participation PSAT	20	25	30	35	40
Percent Participation SAT	25	30	35	40	50
PSAT Verbal	38.5	39.5	40.5	41	42
PSAT Quantitative	37.5	38	38.5	39	40
SAT Verbal	440	450	450	450	450
SAT Quantitative	400	410	420	430	440
10. STUDENT ATTENDANCE					
Average Daily Attendance	86	88	90	92	94
11. TEACHER ATTENDANCE					
Average Days Absent	6	5.5	5	4.5	4
12. SCHOOL CLIMATE AND SAFETY					
Must set own goals based on climate surveys and/or security incidences.					

One problem with systems that rely on absolute goals is often the manner in which such goals are established. In many cases, goals are set based upon what people would like to achieve with no consideration of the probability of making those goals. Educators are often faced with the dilemma of either setting goals too low and being accused of setting low expectations, or establishing goals that are too lofty and having a very slim chance of making them. The issue becomes particularly problematic when part of an individual's evaluation is based upon the degree of goal attainment. The improvement or effectiveness index component of this system, described in the next section of this monograph, can be used to establish meaningful targets based on best practice. This methodology allows the system to provide attainable targets that challenge school staffs but are demonstrably appropriate.

Figure 2. Schematic Depicting the School Improvement Process

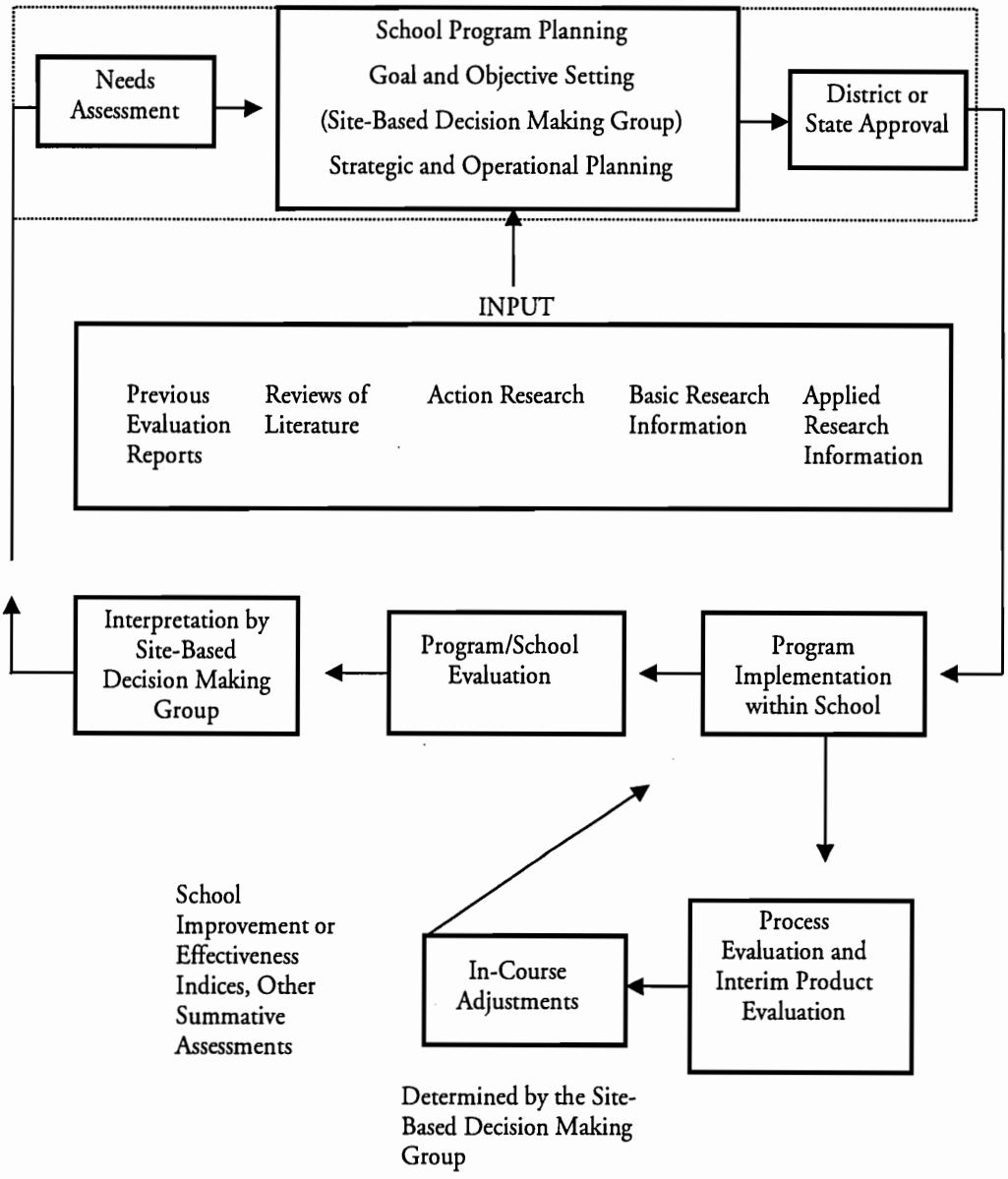


Figure 2 provides a schematic of the school improvement process as it functions within the parameters of site-based decision making. The process begins with each school conducting an annual needs assessment, because a prerequisite to improvement is a knowledge of existing performance levels. Thus, the backbone of any renewal system must be a comprehensive context evaluation program (needs assessment). Context evaluation is the provision of baseline information that delineates the environment of interest, describes desired and actual conditions pertaining to the environment, identifies unmet needs and unused opportunities, and diagnoses the problems that prevent

needs from being met and opportunities from being used. An adequate context evaluation system is founded on a longitudinal data base that is maintained at a district or state level and provides periodic reports on such variables as student dropout, attendance, achievement levels, demographic and vocational patterns, community socioeconomic status and dominant value patterns, and teacher academic and demographic characteristics. Thus, a context evaluation system provides the basis for formulating change objectives by identifying needs and, in some cases, outlining practical constraints in identified problem areas.

Figure 3 outlines the basic components of an operational context evaluation system. In addition to those investigations outlined in Figure 3, which should be conducted annually, a number of specific needs assessments may be conducted as they relate to specific problem areas. Examples might include an assessment of the extent of individualization in school or district classrooms, a survey of drug usage among students, or a study of the perceptions of patrons, educational community, and students re garding the worthiness and effectiveness of current and proposed educational practices. In addition, schools should be encouraged to supplement needs assessment information with local studies, tests, and portfolio assessments.

Figure 3. Examples of Annual Context Evaluation Studies

3.1 *Measurement Profiles*—a summary of the results of the system-wide norm-referenced testing program in addition to community socioeconomic data and a profile of the teaching staff. Results should be aggregated by school community and presented as they relate to national norms. Largely descriptive in nature, these profiles are used to inform educators and patrons of the relative quality of education in the district/region/state and to diagnose gross weaknesses in the instructional program.

3.2 *Criterion-Referenced Testing Profiles*—a summary of the results of criterion-referenced testing programs. These are used as a supplement to the measurement profiles and provide estimates of the functional literacy of school children.

3.3 *Graduate Follow-Up Studies*—a series of studies on graduates of the schools. These include comprehensive data on graduate employment, education, attitudes, life-status, etc. that are used to determine the extent to which educational programs are meeting student needs. The resulting data would then be used as a guide to curriculum planning. One-and-five year follow-ups should be conducted.

3.4 *Dropout Studies*—a series of studies designed to provide descriptive data on dropouts, information about variables associated with dropout, the interactions among such variables, and trends in dropout. Emphasis is on the early identification of potential dropouts so that intervention strategies may be implemented.

3.5 *Input Variable Studies*—a continuous monitoring of the inputs to schools. These studies provide the cost data for cost-benefit and cost-effectiveness analyses, as well as data on the schooling environment.

3.6 *Measurement Studies*—a series of studies on the reliability, validity, and comparability of various test used. These studies provide estimates of the degree of faith that can be placed in test data.

3.7 *Longitudinal Trend Studies*—a series of studies investigating achievement, enrollment, and community trends over time. These studies provide an accountability function.

3.8 *Student Context Studies*—a series of studies designed to determine student course enrollment patterns such as enrollment in honors programs, etc.

3.9 *Teacher and School Effectiveness Indices*—a system designed to produce student-adjusted gain statistics on norm-referenced (fall to fall) and criterion-referenced (spring to spring) tests as well as on other important aspects of instruction. The system identifies teachers and schools that are doing better than expected as well as teachers and schools that are doing more poorly than expected, thus flagging those specific situations for additional study. These value-added indices are discussed in detail later in this monograph.

In order to meet fully the information needs of planning decisions, a context evaluation system must include the capability of providing valid projections of the future level of certain important variables. Figure 4 outlines the general areas of future-oriented projection studies that provide crucial information for most educational decisions.

These studies encompass many variables and are designed to aid decision makers in making intelligent data-based decisions about the future. In addition, projection models dealing with specific problems, such as cafeteria inventory and ordering, could be designed upon request and receipt of high enough priority to allow funding.

Once the context evaluation system has identified needs, site-based decision makers must prioritize those needs and focus upon reducing the discrepancy between desired and existing conditions by establishing goals for those needs that receive highest priority. It is at this point that input evaluation information is brought to bear. Input evaluation is the provision of information for determining methods of resource utilization for accomplishing program goals. In a functioning evaluation system, there are six major sources of input information:

- previous summative product evaluation information, including school effectiveness indices
- review of literature
- basic research information
- applied research information
- action research
- nonresearch and evaluation information.

Figure 4. Examples of Projection Studies

4.1 *Student Demographic and Enrollment Study*—a study designed to project and locate population and provide forecasts of future school enrollments within specified regions for the purpose of providing long-range planning information needed to determine trends and expected demands on educational facilities, staff, and programs.

4.2 *Faculty Flow Study*—a study designed to project the number and cost of teachers required under a multiplicity of policy and/or environmental changes. The study will project the number and characteristics of teachers who will terminate, remain, or need to be hired. Such information is useful for teacher contract evaluation, proposed legislation, evaluation, staffing projections, and hiring/termination analyses.

4.3 *Facilities Study*—a study designed to project the amount, type, and cost of required space areas and to compare projected requirements with the existing inventory of space in order to determine deficiencies or excesses by individual school or demographic area. Such information feeds construction and school attendance zoning decisions.

4.4 *Financial Study*—a study designed to obtain an overall financial projection of district needs based on input from the preceding studies. Features include projection of state-aid funding, debt-service analyses and new bond requirements, revenue and expenditure analyses, and tax-rate-demand analyses.

Summative product evaluation information concerns the extent to which project, program, or school goals are achieved. When product evaluation information is available relative to a given program with goals similar to those identified in response to context evaluation information, that information provides useful input to decision makers in determining the probability that the program would reduce the identified discrepancy.

Basic research information pertains to information about fundamental relationships that affect student learning. Before making a decision to implement a given program, decision makers should be apprised of the extent to which that program is or is not consistent with the principles established by basic research in learning and development. This often requires reviews of the literature.

Applied research information concerns the interaction between student characteristics, teacher characteristics, and instructional systems. Applied research differs from basic research in that the information provided is more closely related to specific decisions in an applied educational setting. Decision makers need information relative to the types of students (e.g., high-anxiety versus low-anxiety) that function best in given instructional systems implemented by teachers with different types of characteristics or traits.

Local school staffs should be encouraged and trained to design, implement, and interpret action research studies. With the movement to site-based management, it is impossible to supply school staffs with sufficient centrally-produced information pertaining to their many and varied needs. Action research is a process for problem-

solving that is designed and implemented at the local building level. It is a process of taking and studying action and its corresponding consequences so that more effective action may be taken (Lewin, 1946; Town, 1973). Expressed sequentially, action research requires a continuous recycling through four steps: 1) identification of needs, 2) development of plans of action to address these needs; 3) execution of these plans of action, and 4) formative evaluation of these plans. In open organizations such as schools, the strength of action research lies in its implementation by the organizations' members in their respective work sites. In effect, members of the organization actively learn while they study problems in contexts that they generally perceive as relevant and important. The results are used to supplement the more formal information available from district or state evaluation departments.

Finally, non-research and evaluation information must enter into most educational decisions. Such information as capabilities of staff members, costs, political feasibility of program implementation, and existing facilities must be taken into account.

After the collection of relevant input information feeding the preliminary program planning stage, school decision makers determine whether or not sufficient resources are available to make the desired instructional changes. Quite often, adequate resources are not available and some compromise is necessary. In many cases, the lack of resources is not limited to the realm of cost and political feasibility but rather stems from an insufficient base of research information. Thus, educators are often in the position of having sufficient material resources but insufficient information resources.

If sufficient material resources are not available, the system may have to exist for some period of time in a state of enlightened persistence. Periodic context evaluation will continue to highlight the extent of discrepancy between that which is desired and that which exists. If the problem results from insufficient information resources, programs are often implemented without sufficient support data and an information base is built through a series of systematic evaluation and applied research studies.

To cope with the problem of insufficient information resources, national development centers should be established and charged with the responsibility of developing instructional systems to meet the needs outlined by context evaluation. Materials and instructional systems are only developed at the local level if no potentially useful materials are available, since the development of instructional systems is an extremely costly proposition.

If sufficient material and information resources are available, or if sufficient material and minimal information resources are available, the extended program planning phase is entered. This is the phase that is entered as a result of the information gleaned from the input evaluation. The evaluator's role in the earlier phase of program planning involved making all relevant, available input information available to program planners. Once it is decided to take a particular course in remediating a demonstrated need, the evaluator must ensure that stated program objectives are measurable.

Out of the program planning sessions, the evaluator develops a detailed program evaluation design specifying the criteria by which the school will be judged. All stakeholders must have input into this design. This design becomes part of the school improvement plan. The development of this evaluation design must necessarily involve continuous interaction between stakeholders and the evaluator in order ultimately to produce maximally effective information. Obviously, the evaluator must be independent of program management to ensure the optimum objectivity of evaluation results.

Once the program implementation phase is entered, the role of the evaluator and the school staff assigned to monitor the implementation of the SIP becomes one of providing continuous formative evaluation reports relative to program implementation. These reports fall primarily into two categories: process evaluation and interim product evaluation. Process evaluation has three major objectives: 1) the detection or prediction of defects in procedural design or its implementation during program implementation stages; 2) the provision of information for programmed decisions; and 3) the maintenance of a record of the implementation procedure as it occurs. Thus, process evaluation information keeps school management informed of the extent to which program implementation conforms to specifications and, from an evaluation standpoint, guards against the evaluation of a fictitious event.

Interim product evaluation provides periodic feedback to school management relative to the attainment of specific subobjectives during the implementation phase. Thus, process and interim product evaluation reports inform program management as to implementation and goal attainment levels while program adjustments are still feasible.

Upon completion of a given cycle of program implementation, a summative product evaluation report is prepared. This report generally addresses the extent to which school objectives were achieved relative to a set of criteria specified in the SIP, as well as the cost-effectiveness of various school implemented programs relative to alternative instructional strategies. Information relative to these areas of concern must be interpreted in light of process and interim product evaluation information. Without information about program implementation, product evaluation information is of little use and it is difficult to chart a course for instructional improvement.

Schools should be encouraged to use portfolios, protocol analysis, and other forms of authentic assessment in monitoring their programs. This information then can be used to provide evidence of accomplishment in instances where the more standard types of assessment fail to show progress. Accomplishing this type of assessment on a districtwide or statewide level is difficult. Performance testing was at one time being built into the Dallas District's Assessment of Course Performance (ACP) test. The ACPs are standard final examinations in seventy-two courses, grades 9-12. One hour was to be multiple-choice while the other hour was to be performance tests. These tests were developed by the evaluation department and had detailed scoring protocols. The performance portion of the tests would have been scored by teachers with random verification of scoring done by the evaluation department. Systemwide performance

testing was subsequently eliminated by district administration as being too time-consuming. While it is not certain that the necessary reliability across scorers and tasks on the performance tests would have been attainable, it is important that the message be communicated to teachers that the kinds of skills and activities measured by performance tests are part of those the district wants them to teach their students. Thus performance testing is more of a curriculum issue than an assessment issue. Early evidence on performance tests suggests that they are much more difficult for students than the average multiple choice tests (Dryden, 1991). Figure 5 (see end of chapter) lists examples of formative and summative data available to schools. These data provide much of the backup data mentioned earlier in reference to Figure 1.

Figure 5 provides information on the indicators, the possible SIP goals that are addressed by each indicator (SIP goals are referenced in Figure 1), the targeted audience(s) for each indicator (1-student, 2-parent, 3-teacher, 4-principal, 5-school community council, 6-central office line staff, 7-central office staff, 8-superintendent, 9-Board of Education and the public), whether the indicator's purpose is primarily accountability or decision making, and when the data should be available.

The reader will note that the majority of the data specified in Figure 5 are for decision making. This is in keeping with the philosophy that accountability information without information for diagnosis and improvement is of little use. In designing an accountability system, it is important to analyze data needs at each level in the organization and to focus data-reporting systems on those needs. One way to accomplish this, particularly in the area of student outcomes, is to identify data needs at the teacher level and then aggregate upward and summarize to meet informational demands at each successive level of the organization. Thus, teachers are provided with timely information necessary to improve instruction and data needs are met at the higher levels of the organization.

Obviously, if school staffs are to use available data effectively to improve schools, a great deal of training must occur. First, school staffs must be taught to collect and interpret data objectively. Second, they must learn to utilize available data in designing and implementing instructional programs. Training modules for school staffs should be developed in keeping and scoring student portfolios of work, designing and scoring performance tests, conducting protocol analysis, developing teacher-made tests, interpreting and using data, and designing and conducting action research.

Second tier: district implementation process

The second tier of the accountability system, the District Improvement Plan (DIP), presents targets and corresponding strategic plans of action with a multi-year planning horizon. The plan meets the accountability objectives and strategic planning requirements of a number of concerned audiences including the general superintendent, the Board of Education, the State Education Agency, school district staff, and the public. The DIP must meet the four major requirements of a strategic planning system in that it receives input from all district departments and campuses, it sets accountability targets and minimum standards of performance for the district and each of its schools,

it provides systemwide plans of action for meeting the major targets of the district, and it specifies the methodology required for monitoring its implementation. DIP targets are in the same areas as the various SIP targets. Each school must do its part to achieve district objectives. If the state is the entity responsible for education, then the DIP would be a state improvement plan.

The DIP contains the strategic plans of the district's or state's support divisions relative to their contributions to meeting the district's or state's targets. It also contains the desired levels of outcomes in the final target year and the intermediate steps necessary to get from baseline levels to desired outcomes. The DIP is directly related to the SIPs in that outcome levels that are specified in each of the SIPs are those levels that will help the district reach its goals. The DIP sets the criterion level for desired outcomes. Goals are absolute, but should be specific to each school. All schools could meet them or no schools could meet them; that is, target accomplishment is not determined by a norm group. DIP targets could also be established empirically based upon best practice.

Third tier: school effectiveness indices

The final tier of the accountability system is the most important from the standpoint of evaluating schools. Inherent in the task of evaluating schools are two complex issues: how to define effectiveness, and how to develop a model to assess effectiveness.

In an attempt to provide a better definition of effectiveness and respond to the narrowly-focused concern of earlier effective schools research, Murnane (1991), David (1987), and others have been proponents for developing an expanded number of outcome indicators. In addition, Oakes (1989), David (1987), and Cohen (1986) have argued the importance of incorporating input and process/context indicators as important aspects of better accountability mechanisms.

Possible input indicators often include school enrollment, socioeconomic/ethnic composition, proportion of limited-English-speaking children, enrollments in categorical programs, staff characteristics, and financial resources. Process indicators describe what is being taught, the way it is being taught, and include consensus on school goals, instructional leadership, opportunity to learn, school climate, staff development, and collegial interaction among teachers. Outcome indicators are usually related to identifying the effects of school on students or providing information about other definitions of "good schooling," and may include student academic performance, teacher and student attendance rates, dropout and completion rates, performance of students at the next level of schooling, parent and student satisfaction, percent completing advanced courses, college attendance, and individual school goals (David, 1987; Oakes, 1989; Olson and Webster, 1986; Pollard, 1987; Shavelson et al., 1987).

The anatomy of effectiveness indices

The school effectiveness methodology that is being proposed here defines a school's effectiveness as being associated with exceptional measured performance above or below

that which would be expected across the entire reference group (which could be district, state, or nation). When a school's population of students departs markedly from its own preestablished trend or from the more general trend of similar students throughout the reference group, this departure is attributed to school effect. The problem of measuring a school's effect, then, becomes one of establishing the student levels of accomplishment on the various important outcome variables, setting levels of performance based on these predictions, and determining the extent to which its students, on the average, exceed or fall short of expectation. The statistics procedures for measuring a school's effect involve the utilization of HLM for student-level variables and multiple regression analysis for school-level variables to compute prediction equations by grade level for each outcome variable independent of school identification and then using those equations with students or schools to obtain gains over expectations. Relative weights could be assigned to the outcomes by an accountability task force or some other form of stakeholder group. Once weighted levels of performance have been determined, the methodology provides an indicator of how well a school performs relative to other schools throughout the reference group. To a great extent, the same targets used in the SIP and DIP processes would be used as outcome variables in the school effectiveness indices. Thus, schools work on improving target variables in an absolute sense through their SIPs and are judged in terms of goal attainment, improvement, and effectiveness.

One approach to developing a value-added school evaluation model is OLS regression. The basic OLS regression model is generated from the standard OLS equation. This is represented by equation (1) for student-level variables:

$$(1) \qquad Y_1 = \beta_0 + \beta_1 X_i + r_i \qquad \text{where } r_i \sim N(0,\sigma^2)$$

Using this model, the Y represents any of the outcome variables in the system. The X represents a predictor variable available for the model in question. (These values are without reference to school at the moment.) After a solution is found for X, the model is solved for each student and the value of the residual r_i is determined. This value of r represents the value-added portion of the student's score plus any individual error for the student on the particular outcome measure (Y). This equation is solved for each of the possible Y variables and the student residuals determined for each student and variable. As predictors are added, including more prior or concomitant variables, the standard OLS multiple regression equations are used and solved in a similar fashion.

Once student residuals are obtained, the relative effectiveness for each school is determined through the following steps:

1. The values of the r_i are grouped by each school.
2. The residuals are summed and a mean residual is determined for each school.
3. The mean residuals are corrected for either sample size or for shrinkage, which adjusts the means for sample size and differential variation.
4. The means for each of the individual Y outcomes are standardized to a unit scale.
5. The means for each Y are weighted if the variables are differentially weighted and are summed across the Y variables.

6. These weighted means are standardized and rescaled to the final scale chosen for the effectiveness scores.

In the briefest manner, these steps represent the general outline of an OLS regression-based value-added school evaluation model. Note that several steps are indispensable to this process. Step 3 is critical. If the means are not adjusted for either sample size or for shrinkage, the resulting means will have different variances. In this case, means from schools with small sample sizes will be biased away from the district mean. These schools will have an *a priori* probability of being judged more or less effective than larger schools when no other differences exist.

The second critical step is 4. The means across variables must be standardized to the same mean and standard deviation before they are weighted and summed. Failing to do so will put undue emphasis on variables with larger variances.

A more detailed discussion of the OLS regression model and its application can be found in Webster et al. (1994, 1995, 1996, 1997a). These papers also carefully explain the application of the OLS model as it is applied in the Dallas Public Schools' value-added accountability system. The Dallas system will be discussed in an overall fashion later in this paper, but these papers have more complete explanations.

Within this general framework for an OLS regression model, many variations are possible. However, the reader must be aware of what each variation entails and how it will affect the results. Before describing the different approaches, a point discussed earlier should also be kept in mind: all of these approaches are a significant improvement on comparing schools with unadjusted outcomes. If an approach does not control some variables, that simply means a better value-added approach is possible. Further, when we note that a variable is not controlled sufficiently, it generally means sufficient relative to another value-added model. It generally does not mean the level of control is insufficient in general. In many instances, where two values are described as one better and one worse, both are generally acceptable and the difference is minute. That noted, the general types of OLS models are as follows:

1. The predictor is limited to one prior measure of the outcome variable. This is the simplest situation. For example, reading in the prior year is used to predict reading in the current year and no other variables are included. Problems with the model are that student-level concomitant variables are inadequately controlled, school-level concomitant variables are inadequately controlled, and individual student predictor scores are subject to more anomalous influences, one of the greatest being cheating on the test (Webster et al., 1995). On the positive side, any model with only one year of prior information preserves the largest number of subjects in the analysis. We have found that each additional year of prior data included in the model results in approximately 8 to 10 percent missing data. This model can also be used for school-level outcome variables. In this case, steps 1-3 are obviously superfluous, although a weighted regression based on school size can be employed.

2. Two or more years of prior variables are used and no concomitant variables are included. This is a longitudinal model. Again student and school concomitant variables are inadequately controlled. Missing data becomes a problem (Webster et al., 1997a). On the positive side, anomalous influences are better controlled. This model, with two years of predictors, is the model we recommend for school-level outcome variables.

3. One or more years of prior variables are used and student-level concomitant variables are used. This model controls the correlations of the residuals with student-level characteristics very well. The correlation with school-level characteristics is improved but still not controlled adequately (Webster et al., 1995).

4. One or more years of prior variables are used and student and school-level concomitant variables are used. This model controls student-level characteristics well and offers improved control of school-level variables, but not complete control (Webster et al., 1997a).

In studying the various HLM value-added models, our research has been limited to the Bryk and Raudenbush (1992) conception of the model. Not having specifically tested the versions of the model that are used by other researchers in other situations, we will strictly limit our remarks to the HLM approach we have tested and we use in our own value-added model. However, we see no underlying theoretical reasons why the results we have accumulated should not generally be applicable to other formulations of the mixed-model methodologies.

The standard equations for the random effects HLM model are given in equations 2 through 4 for a single level 1 predictor and a single level 2 conditioning variable. Note that level 1 contains a model of school-level data. The two types of data are modeled simultaneously in an HLM model. The significance of this point will be brought out in the discussion of the model. As in the case of the OLS regression model, these equations can be expanded by the inclusion of more level 1 student predictor variables (X) and the inclusion of more level 2 school conditioning variables (W). School effects are estimated directly from shrinkage-adjusted empirical Bayes residuals resulting from the application of the HLM model (Bryk and Raudenbush, 1992). Again, our research papers contain more explicit formulations of the model under many different conditions. The interested reader is referred to Webster et al. (1995), Mendro et al. (1995), Orsak et al. (1997), Weerasinghe et al. (1997), or Webster et al. (1996, 1997a) for more detailed models and discussions of these applications.

(2) Level 1 $$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij}$$

(3) Level 2 $$\beta_{0j} = \Upsilon_{00} + \Upsilon_{01}W_j + u_{oj}$$

(4) $$\beta_{1j} = \Upsilon_{10} + \Upsilon_{11}W_j + u_{1j}$$

where $r_{1j} \sim N(0,\sigma^2)$
$u_{0j} \sim N(0,\tau_{00})$
$u_{1j} \sim N(0,\tau_{11})$
$Cov(u_{0j}, u_{0j'}) = \tau_{01} = \tau_{10}$

As in the situation with the OLS model, there are many choices within an HLM model. However, we will not explicate them other than in the discussion of our research with the model. We feel that HLM provides a significant improvement in controlling school-level concomitant variables and that its use should be concentrated on that function.

Generally, then we note the following:

1. The HLM model should be used as a random-effects model, as noted in the equations 2 through 4. It is possible to use the HLM model as a fixed-effects model which allows the inclusion of a large number of student-level variables at level 1, but this produces possible suppression effects among the student-level variables (Webster, et. al., 1997a).

2. The use of HLM as a random model provides the best approach to control specific school-level variables and to provide a control for unspecified student-level variables (Webster, et al., 1997a). It also requires that a modified model be used to include many student-level variables, since the random HLM model will not allow a large number of such variables directly (Webster, et al., 1996, 1997a).

3. Use of HLM without school-level conditioning variables at level 2 provides no significant benefits over an OLS model in value-added school effectiveness models (Mendro, et al., 1995; Weerasinghe, et al., 1997; Webster, et al., 1995, 1996, 1997a).

4. There is no significant difference between HLM models based on predicting an outcome from a predictor variable and models based on gain scores when pretest is used as a predictor (Weerasinghe, et al., 1997). Without pretest as a predictor, gain models are significantly negatively correlated with pretest.

THE DALLAS VALUE-ADDED ACCOUNTABILITY MODEL

Figure 6 (see end of chapter) displays the variables used to compute the school effectiveness equations used in the Dallas Public Schools. Each outcome variable is described under "outcome" along with the grades at which it is included, the score that is the basis for the analysis, the methodology utilized, the level at which the data are analyzed (student or school level), possible predictors and the grades at which they are found, and possible school-level conditioning variables included in the student-level equations. Two different regression models are used depending on whether the unit of analysis is the student, in which case hierarchical linear modeling is used, or the school, in which case multiple regression analysis is used. Through these approaches it is possible to obtain extremely reliable predictions of student and school outcomes and to compare actual to predicted outcomes. All analyses that are done at the student level are calculated on residuals, that is, statistics that have had individual student characteristics over which the schools have no control removed from the equations (gender, ethnicity, limited English-proficient status, socioeconomic status, and all of the interactions between those variables).

Using the HLM model as a random effects model requires some modification in order to include a broad array of student-level variables, as noted above. The Dallas value-added model makes these modifications for both mathematical and political reasons (Webster and Mendro, 1995). The intent of using a broad array of student-level concomitant variables is to eliminate specific effects due to these differences. There are always skeptics among educators and the public who feel that this is not possible. In partial response to this skepticism, early in the process of developing the Dallas value-added model, we adopted a two-stage regression system.

In this system, the student-level concomitant variables described in Figure 6 are regressed against both the outcome variables and the prior predictor measures. The residuals from these regressions are then used in a random effects HLM system with an array of school-level conditioning variables. From a mathematical standpoint, without this two-stage process, a one-stage fixed-effects HLM model would be required. This is not our preferred model for the reasons noted, specifically the possibility of suppressor effects. From a political standpoint, the mean residuals for the predictors and outcome variables can be computed by lunch status, gender, and ethnicity/language proficiency categories and can be shown to be equal. Being able to show that the effects of these variables are controlled eliminates a large amount of skepticism, although not all (Webster et al., 1997a).

A second procedure specific to the Dallas model is the adjustment of residuals in the first stage of the two-stage model. After residuals are computed in the first stage, the predictor space is divided into 256 equal intervals, and the residuals within each interval are standardized to a mean of 0 and standard deviation of 1. This process assures that differing residual variance and means across the predictor space will not affect the value-added estimates (Mendro, et al., 1995; Weerasinghe et al., 1997; Webster et al., 1995, 1996, 1997a).

To summarize the Dallas value-added accountability model from the discussion in this entire section:

1. School variables are predicted in a regular OLS regression using two years of prior outcome variable data. Effectiveness scores are computed from the residuals of the regression.

2. Student variables are predicted from a two-stage, modified OLS regression and HLM regression.

3. The first stage of the student variable process regresses outcome variables and prior predictor variables against student-level concomitant variables, adjusts the residuals for homogeneity, and provides residuals for the HLM stage.

4. The second stage of the student variable process uses one year of prior level residuals from the first stage to predict the outcome residuals from the first stage in a two-level

HLM random effects model with an array of school-level conditioning variables at the second-level.

5. The results of each HLM analysis by student outcome variable and the school-level outcome variable OLS regressions are standardized and weighted by Accountability Task Force-determined weights.

6. The weighted results are combined to give a total school effectiveness estimate for each school.

Additional benefits of the Dallas model (many of which would accrue from other models using the strictures discussed earlier) include the following:

1. The system gives individual school rankings on each outcome variable. This allows the construction of school profiles on a variable-by-variable basis.

2. The student residuals from the HLM analyses can be isolated and regrouped by the lunch, gender, and ethnicity/language proficiency categories by outcome variable, which allows schools to determine whether their educational efforts are biased for or against a particular student group.

3. The testing percentage requirements have resulted in high levels of Dallas students participating in the testing on all outcome measures.

4. The use of many outcome measures has reduced some of the tendency of schools to concentrate on only one test or outcome. (By no means has it eliminated the tendency of some schools to concentrate on the state accreditation test, but the ability to profile each outcome variable makes it easy to identify which schools have concentrated too heavily on a particular variable.)

USES OF VALUE-ADDED SYSTEMS IN SCHOOL EVALUATION

The results from value-added accountability systems have many uses in school evaluation. This comes from the characteristic of these systems that student data are analyzed at the individual student level and residuals are available by student. Thus, wherever the student composition of a group, a class, a program, or a project are known, value-added effectiveness estimates can be constructed for these groupings.

Sometimes the groupings are conveniently available as schools. For example, an experimental mathematics program involving all the fourth-grade students at three schools allows for the easy comparison of the value-added estimates for mathematics outcomes for the three schools. Because of the nature of these estimates, the entire district forms a reference point for the comparison of the specific outcomes. Hence, if all three are above the district in effectiveness, it adds evidence to the assertion that the program is having an effect. Combined with appropriate process evaluation and the results of

measures not included in the outcome measures (specific criterion-referenced tests, for example), value-added measures add a considerable weight to the evaluation. In particular, efforts involving entire schools are easily measured using the value-added outcome for the whole school. Thus, for example, the restructuring of entire schools, the selection of schools to participate in school-community participation programs, or any other schoolwide efforts which purport to affect the outcome measures can easily have a potent dimension added to the evaluation through the use of the schoolwide, value-added measure.

Where groups are available by program, the collection of results by outcome variable and program provides a quick reference. For example, a program involving selected students in a school can readily be evaluated by taking the student residuals for the program and non-program students and comparing them. Again, all scores are referenced to the district as a whole and the value-added estimates for program and non-program students give an unbiased look at the relative merits of the program.

By using a model similar to the one described above, the influence of important student and school-level contextual variables, over which the schools have no control, are eliminated from the equations. Schools derive no particular advantage by starting with white or minority students, rich or poor students, LEP or English-proficient students, or male or female students. Schools with large concentrations of various combinations of students are also neither advantaged or disadvantaged. Other variables such as mobility and crowding are also controlled at the school level. Research on these models has shown that the results produced do not significantly correlate with any of the individual student or school-level contextual variables (Webster et al., 1995, 1996, 1997a). Thus, the "playing field is level" and practitioners' concerns about the impact of background variables on measured school effect is allayed.

In addition, because of the value-added nature of the equations, schools derive no particular advantage by starting with high or low-scoring students. Equations set individual predictions for each student based on that student's placement on the pretest(s) of interest. Lower-scoring students have lower predicted scores. Higher-scoring students have higher predicted scores. Equations must be developed at the individual student level, not the school level, to accomplish this.

At this point, the reader may ask whether the added precision gained from a regression-based value-added model is worth the added trouble. Why not just evaluate schools based on absolute test scores or on unadjusted gain scores? The following case study of a system that did just that is illustrative of the problems that one encounters and the lack of fairness inherent in such a system.

In this case, a major state education agency, like many other state education agencies in the United States, developed its own accountability system. This system was largely limited to results of a state criterion-referenced test, student attendance, and dropout rate. The system reports data on districts in both a cross-sectional and cross-sectionally longitudinal manner and purports to allow comparisons of districts to "like" districts

across the state as well as to the state as a whole. No cohorts of students are used. Although we do not intend to provide a thorough critique of this accountability system, the system provides excellent examples of inappropriate methodology and interpretation.

The primary measure on which the system is based is a criterion-referenced test. Serious questions have been raised as to the reliability, validity, and scaling of that test for the purposes for which it is used. There is also the often-quoted concern that relates to basing an entire accountability system on one test. An insistence on not releasing current information on the test cloaked the testing program in a veil of secrecy and only added to the uninterpretability of results and the discomfort of users. (Sample tests are now released after a court confrontation.)

Even overlooking the possible flaws in the test, there are still a number of difficulties with the accountability system. First, it is based upon arbitrary goals: that is, goals that have not been empirically established and have not taken into consideration the difficulty levels or characteristics of the tests. Second, the first phase of the system is based upon unadjusted test scores. This means that students, not schools, are being evaluated. As previously mentioned, the technique of comparing schools based upon unadjusted outcome measures adversely affects schools with student demographics that differ from the norm. This is particularly true of schools with large minority and poor student populations.

To further illustrate this point, Table 4 displays some of the demographic characteristics of the top 20 percent of schools as defined by the previously-described school effectiveness methodology and compares those schools to the state rank. The reader will note that effective schools, as defined by this methodology, come in all sizes and shapes. District statistics at the particular grade levels are also presented to provide a framework for interpretation of the information.

At the K-6 level, the most effective schools tended to have smaller enrollments than the average enrollment of district elementary schools. Enrollments ranged from a low of 193 to a high of 860. Ethnicities ranged from a high of 99.7 percent Black, 90.4 percent Hispanic, and 64.5 percent White to a low of 3.5 percent Black, 0.3 percent Hispanic, and 0 percent White. Most deprivation indices were above the district average of 69, ranking as high as 92, while the percentage of limited-English-proficient students ranged from a high of 57.9 percent to 0. In short, whether or not a school was ranked among the most effective could not be predicted from the demographics of the students that it served. This is because the system is based on improvement, not upon absolute levels of achievement and associated variables.

The last column in Table 4 (state rank) depicts the school rank based on the percentage of students passing all subtests of the state criterion-referenced test. The top twenty-seven K-6 schools in the district on the effectiveness indices had composite ranks between 3 and 107 when ranked based on absolute achievement levels. It should be noted that the six schools that ranked in the top fifteen in the district on the state

Table 4. Demographic Characteristics of the Top Twenty Percent of Effective Schools

Rank	Grades	Enrollment	Percent White	Percent Black	Percent Hispanic	Percent DEP	Percent LEP	State Rank
1	K-3	555	0.4	98.9	0.7	87	0.1	20
2	K-6	238	4.5	15.3	79.7	84	57.9	94
3	K-3	447	2.5	80.1	17.1	84	8.3	26
4	K-3	194	0	98.0	1.5	77	1.5	58
5	K-3	573	0.9	57.4	41.7	80	36.5	77
6	K-6	529	0.2	96.4	3.5	92	2.6	39
7	4-6	193	0.5	85.3	12.7	75	5.7	60
8	4-6	336	1.5	64.0	34.2	87	17.6	83
9	K-6	518	64.5	18.5	10.2	20	2.7	11
10	K-6	462	54.4	16.2	28.4	33	3.5	4
11	4-6	398	0.8	75.5	23.5	71	15.3	88
12	K-6	539	51.9	11.0	31.4	40	15.6	8
13	K-6	656	50.0	27.5	21.7	36	13.4	9
14	K-6	830	37.0	37.8	23.4	51	13.1	26
15	K-6	776	0	99.7	0.3	75	0	50
16	K-3	214	0.5	87.3	12.2	64	8.4	107
17	K-6	630	63.3	7.8	26.0	27	15.3	12
18	K-6	680	0.2	99.2	0.6	51	0	32
19	K-6	569	0.2	99.0	0.9	85	0	30
20	K-6	741	58.6	11.0	20.8	36	12.0	17
21	K-6	860	0.2	88.8	9.6	93	5.0	71
22	K-6	702	4.4	3.5	90.4	81	52.1	86
23	K-6	697	39.0	27.5	29.8	47	22.7	17
24	K-6	571	3.0	20.5	76.1	92	53.9	91
25	K-6	483	55.4	11.3	30.9	18	3.5	3
26	K-6	382	41.1	37.1	19.5	21	0	39
27	K-6	331	0	99.7	0.3	64	0	62
District	K-6	592	16.1	43.7	38.2	69	23.2	

Rank	Grades	Enrollment	Percent White	Percent Black	Percent Hispanic	Percent DEP	Percent LEP	State Rank
1	7-8	693	13.6	30.8	52.4	85	31.2	10
2	7-8	668	30.3	37.6	29.7	45	15.4	6
3	7-8	888	18.2	16.3	63.2	64	29.9	9
4	7-8*	367	24.4	50.4	22.4	38	0	2
5	7-8	863	7.0	75.4	15.7	30	1.3	5
District	7-8	703	15.2	48.6	34.3	55	13.2	

Rank	Grades	Enrollment	Percent White	Percent Black	Percent Hispanic	Percent DEP	Percent LEP	State Rank
1	9-12	1129	0.4	97.9	1.7	32	0	16
2	9-12	1004	36.7	37.7	23.3	24	13.0	4
3	9-12*	3567	18.5	43.7	32.9	22	4.8	5
4	9-12*	129	46.9	31.5	17.7	9	0	1
5	9-12*	623	48.4	34.3	15.6	9	0	2
6	9-12*	644	10.2	60.0	25.5	27	1.7	8
District	9-12	1093	15.8	49.8	31.5	28	10.1	

system, when no known non-school sources of variation were accounted for, were at least 50 percent White and had no deprivation index above 40. On the state level, the highest rankings on this system are generally awarded to schools serving high percentage of White, economically-advantaged students.

At the 7-8 level, the most effective schools had enrollments varying from 367 to 888 and were from 7.0 to 30.3 percent White, 16.3 to 75.4 percent Black, and 15.7 to 63.2 percent Hispanic. Deprivation indices varied from a low of 30 to a high of 85 percent and limited English proficient varied from 0 to 31.2 percent. Ranks of the top five middle schools varied from 2 to 10 on the TAAS. These ranks were somewhat closer because Dallas middle schools do not vary as much in demographics as do K-6 schools, and there are not as many middle schools.

At the 9-12 level, magnet schools dominated the rankings. Four of the top six high schools were magnet schools. This finding was predictable, because at this level, magnet school students tended to be more motivated because they had to overtly choose their schools. However, the most effective high school in the district, a school that is 97.9 percent Black and had a deprivation index of 32, was only ranked sixteenth out of twenty-six high schools on the state system.

School effectiveness indices are about fairness. Schools have no control over the students that they receive but should be held accountable for educating effectively the students that they do receive. Effective schools are good schools. Good schools are schools that effectively improve their student's performance on measures that matter.

UTILIZATION OF VALUE-ADDED INFORMATION

The Dallas value-added accountability model has been utilized since 1991-92 to provide unbiased measures of school effectiveness. Since its inception, more detailed information from the system has also been computed and given to schools. The information currently provided includes:

- A total-weighted measure of school effectiveness that provides the school, the administration, the Board, and the public with the answer to the question "What is the overall effectiveness of this school when considering all of our valued outcomes?"

- Measures by grade and outcome variable of each school's effectiveness with:
  –A sorting of measures by grade level to aid in determining the effectiveness of each grade-level team.
  –A sorting of measures by major subject matter category (language arts, mathematics, science, and social studies) to aid in determining the effectiveness of each core curriculum program in the school.
  –Graphical representations of effectiveness by grade to give less numerically-oriented administrators an efficient summary of outcome achievement.
  –Distributions of school effectiveness at each grade for each outcome variable, disaggregated by ethnicity, language proficiency, gender, and socioeconomic status designed to show whether groups are being fairly served by the school.

Added to this, the district's Division of Research, Evaluation, and Information Systems provides regular training, both scheduled and upon request, to administrators, the Board, subdistricts, schools, parent groups, the Accountability Task Force, and to outside organizations on the indices in general, on using the results in overall and specific school evaluation and on using the results for general and specific planning (Bearden, 1997).

These feedback measures from the SEI and their component information are used regularly in school and individual planning and appraisal. The major school and personnel evaluation systems of the district have been coordinated and the results are used in teacher appraisal (Bembry, 1997), principal and school administrator evaluation, and school improvement planning (Webster 1997b). In each instance, a three-step appraisal/evaluation model is employed. In the first step, all data sources are analyzed and a needs analysis is prepared, with the proviso that it must include value-added results where they are available. In the second step, documentable and specific strategies and remedies are devised and specified that the teacher, principal, and/or school is responsible for implementing. These strategies are then implemented during the year with interim checking, assessment, and revision, if necessary. The third step is the evaluation of the accomplishment of the strategies using documentation specified in step 2 and collected during the school year. The advantages of having all of these systems aligned are that they are all data-driven with the value-added measures forming the most important (but by no means, the only) data source; that all systems are focused on the primary *raison d'être* of a school, helping students learn; and that all individual and school evaluations and plans keep improved instruction as their only element.

The implementation of the School Effectiveness Indices model has corresponded with a period of generally increasing achievement, reduced dropout rates, increased attendance, and other improved measures directly addressed by the accountability system. The indices form the basis for a six-year-old program of acclamation and monetary rewards for effective schools and their staff members. The indices are used regularly by the administration as a primary measure for the selection and retention of principals. The indices have been used as the basis of studies of effective and ineffective schools and their differences. The indices also were the major measure in the selection of schools when the decision to restructure a number of schools was made (Webster, 1997b). The studies of school effectiveness using the indices have resulted in better pictures of effective schools, their climates and environments, and the roles played by principal and teachers that influence the degree to which these schools are effective. Finally, as this monograph notes, the school and student information from the School Effectiveness Indices are used regularly in the District's program evaluation efforts.

REFERENCES

Aiken, L.S. and West, S.G. (1991). *Multiple Regression: Testing and Interpreting Interactions.* Newberry Park, CA: Sage.

Bano, S.M. (1985). *The Logic of Teacher Incentives.* Washington, DC: National Association of State Boards of Education.

Bearden, D.K. (1997). *An Overview of the Elementary Mathematics Program Evaluation.* Report REIS97-116-2. Dallas: Dallas Public Schools. (In progress.)

Bembry, K. (1997). *Implementation of the Dallas Public Schools Teacher Evaluation System: 1996-97.* Report REIS97-120-2. Dallas: Dallas Public Schools.

Bock, R.D. (1989). *Multilevel Analysis of Educational Data.* San Diego: Academic Press.

Bryk, A.S., Raudenbush, S.W., Seltzer, M., and Congdon, R. (1988b). *An Introduction to HLM: Computer Program User's Guide* (2nd ed). Chicago: University of Chicago.

Bryk, A.S. and Raudenbush, S.W. (1992). *Hierarchical Linear Model: Applications and Data Analysis Methods.* Newberry Park, CA: Sage.

Bryk, A.S., Raudenbush, S.W., Seltzer, M., and Congdon, R. (1988a). *Toward A More Appropriate Conceptualization of Research on School Effects: A Three-Level Hierarchical Linear Model.* Multilevel Analysis of Educational Data. San Diego: Academic Press.

Campbell, D.T. and Stanley, J.C. (1963). "Experimental and Quasi-Experimental Designs for Research on Teaching." In N.L. Gage (ed). *Handbook of Research on Training.* Chicago: Rand McNally.

Cohen, J. (1968). "Multiple Regression as a General Data-Analytic System." *Psychological Bulletin,* 70, 426-443.

Cohen, J. and Cohen, P.(1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Laurence Erlbaum.

Cohen, M. (1986). Designing State Education Assessment Systems. *Study Group on the National Study of Student Achievement.*

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McParland, J., Mood, A.M., Weinfeld, F.D., and York, R.L. (1966). *Equality of Educational Opportunity.* Washington, DC: U.S. Department of Health, Education, and Welfare, Office of Education.

Comer, J.P. (1988). "Educating Poor Minority Children." *Scientific American,* 259, 5: 42-48.

Cook, D.L. (1966). "Program Evaluation and Review Techniques, Applications in Education." *U.S. Office of Education Cooperative Monograph.* 17, (OE-12024).

Council of Chief State School Officers (1995). *State Education Accountability Reports and Indicator Reports: Status of Reports Across the States.* Washington, DC: CCSSO.

Cronbach, L.J. (1963). "Course Improvement Through Evaluation." *Teachers College Record,* 64, 672-683.

Dallas Public Schools (1996). *School Performance Improvement Awards, 1996-97.* Dallas: DISD.

Darlington, R.B. (1990). *Regression and Linear Models.* New York: McGraw-Hill.

David, J. (1987). *Improving Education with Locally Developed Indicators.* New Brunswick: Center for Policy Research in Education, Rutgers University.

Dempster, A.P., Rubin, D.B., and Tsutakawa, R.V. (1981). "Estimation in Covariance Components Models." *Journal of the American Statistical Association,* 76, 341-353.

Dryden, M.E. (1991). *Evaluation of the 1990-91 South and West Dallas Learning Centers.* Dallas: Division of Evaluation and Planning Services, DISD.

Edwards, M.E. (1991). *School-Centered Education: A Plan for Creating a System of Schools in the DISD.* Dallas: DISD.

Eisner, E. (1976). "Educational Connoisseurship and Criticism: Their Form and Functions in Educational Evaluation." *Journal of Aesthetic Education*, 3-4, 10: 135-150.

Elston, R.C. and Grizzle, J.E. (1962). "Estimation of Time Response Curves and Their Confidence Bands." *Biometrics*, 18, 148-159.

Felter, M., and Carlson, D. (1985). "Identification of Exemplary Schools on a Large Scale." In Austin and Gerber (eds). *Research on Exemplary Schools.* (pp. 83-96). New York: Academic Press.

Gallegos, A. (1994). "Meta-Evaluation of School Evaluation Models." *Studies in Educational Evaluation*, 20, 41-54.

Glass, G.V. (1978). "Standards and Criteria." *Journal of Educational Measurement*, 15, 237-261.

Goldstein, H. (1987). *Multilevel Models in Educational and Social Research.* New York: Oxford University Press.

Henderson, C.R. (1984). *Applications of Linear Models in Animal Breeding.* Guelph, Canada: University of Guelph.

Jaeger, R.M. (1992). "Weak Measurement Serving Presumptive Policy." *Kappan*, 74, 2: 118-128.

Joint Committee on Standards for Educational Evaluation. (1994). *The Program Evaluation Standards.* Kalamazoo, MI: The Evaluation Center, Western Michigan University.

Kirst, M. (1986). "New Directions for State Education Data Systems." *Education and Urban Society*, 18, 2: 343-357.

Klitgaard, R.E., and Hall, G.R. (1973). *A Statistical Search for Unusually Effective Schools.* Santa Monica, CA: Rand Corporation.

Laird, N.M. and Ware, H. (1982). "Random-Effects Models for Longitudinal Data." *Biometrics*, 38-963-974.

Lessinger, L.M. (1970). *Every Kid a Winner: Accountability in Education.* New York: Simon and Schuster.

Lewin, K. (1946). "Action Research and Minority Problems." *Journal of Social Issues*, 2, 58-73.

Lindquist, E.F. (ed). (1951). *Educational Measurement.* Washington, DC: American Council on Education.

Longford, N.T. (1987). "A Fast Scoring Algorithm for Maximum Likelihood Estimation in Unbalanced Mixed Models with Nested Random Effects." *Biometrika*, 74, 4: 817-827.

MacKenzie, D. (1983). "School Effectiveness Research: A Synthesis and Assessment." In P. Duttweiler (ed). *Educational Productivity and School Effectiveness.* Austin, TX: Southwest Educational Development Laboratory.

Madaus, George F., Scriven, M.S., and Stufflebeam, D.L. (1983). *Evaluation Models: Viewpoints on Educational and Human Services Evaluation.* Hingham, MA: Kluwer Academic Publishers.

Mason, W.M., Wong, G.Y., and Entwistle, B. (1984). "Contextual Analysis Through the Multilevel Linear Model." In Leinhardt (ed). *Sociological Methodology.* (pp. 72-103). San Francisco: Jossey-Bass.

May, J. (1990). *Real World Considerations in the Development of an Effective School Incentive Program.* ED 320 271.

Millman, J. (ed). (1981). *Handbook of Teacher Evaluation.* Beverly Hills, CA: SAGE.

Murnane, R.S. (1991). "The Case for Performance-based Licensing." *Kappan,* 73-2, 137-142.

Oakes, J. (1989). "What Education Indicators? The Case for Assessing the School Context." *Educational Evaluation and Policy Analysis,* 11, 181-199.

Olson, G.H. and Webster, W.J. (1986). *Measuring School Effectiveness: A Three-year Study.* San Francisco:American Educational Research Association.

Pollard, J. (1987). *Viewpoints from Selected States on Accreditation and Accountability.* Austin, TX: Southwest Educational Development Laboratory.

Provus, M. (1971). *Disciplinary Evaluation.* New York: McCruthan.

Rosenberg, B. (1973). "Linear Regression With Randomly Dispersed Parameters." *Biometrika,* 60, 61-75.

Saka, T. (1989). *Indicators of School Effectiveness: Which are the Most Valid and What Impacts Upon Them?* American Educational Research Association, San Francisco. ERIC ED306277.

Sanders, W.L. and Horn, S.P. (1995). "The Tennessee Value-Added Assessment System (TVAAS): Mixed Model Methodology in Educational Assessment." In A.J.Shinkinfred and D.L.Stufflebeam. *Teacher Evaluation: Guide to Effective Practice.* Boston: Kluwer.

Scriven, M.S. (1967) "The Methodology of Evaluation." In R.E. Stake (ed). *Curriculum Evaluation.* AERA Monograph Series on Curriculum Evaluation (Vol. 1). Chicago: Rand McNally, 1967.

Shavelson, R.J., McDonnell, L.M., Oakes, J., and Carey, W. (1987). *Indicator Systems for Monitoring Mathematics and Science Education.* Santa Monica, CA: Rand Corporation.

Stake, R.E. (1967). "The Countenance of Educational Evaluation." *Teachers College Record.* 68, 523-540.

Stufflebeam, D.L. (1966). "A Depth Study of the Evaluation Requirement." *Theory into Practice.* 5, (June), 121-134.

Stufflebeam, D.L. (1996). "Evaluating School Districts, Programs, and Personnel: Toward a Unified Approach." Paper presented at the 5th Annual National Evaluation Institute. Bethesda, Maryland, July 1996.

Stufflebeam, D.L. (1967). "The Use of and Abuse of Evaluation in Title III." *Theory into Practice.* 6, (June), 126-133.

Stufflebeam, D.L. and Webster, W.J. (1980). "An Analysis of Alternative Approaches to Evaluation." *Educational Evaluation and Policy Analysis.* 3, 2: 5-19.

Stufflebeam, D.L. et al. (1971). *Educational Evaluation and Decision-Making.* Itaska, IL: Peacock.

Town, S.W. (1973). "Action Research and Social Policy." *Sociological Review* 12, 4: 128-137.

Tyler, R.W. (1949). *Basic Principals of Curriculum and Instruction.* Chicago: University of Chicago Press.

Webster, W.J. (1988). "The Practice of Evaluation in the Public School." In T.C. Dunham (ed). *The Practice of Evaluation*, pp. 24-25. Minneapolis: University of Minnesota Press.

Webster, W.J. (1997b). *Rewarding Effective Schools—Theory and Practice in an Outstanding Schools Awards Program.* Chicago, Illinois: AERA.

Webster, W.J., Mendro, R.L., and Almaguer, T. (1994). Effectiveness Indices: A Value Added Approach to Measuring School Effect. *Studies in Educational Evaluation*, 20, 113-145.

Webster, W.J., Mendro, R.L., Bembry, K., and Orsak, T.H. (1995). *Alternative Methodologies for Identifying Effective Schools.* Distinguished Paper Session, American Educational Research Association, San Francisco. ERIC EA 027 189.

Webster, W.J., Mendro, R.L., Orsak, T.H., and Weerasinghe, D. (1996). *The Applicability of Selected Regression and Hierarchical Linear Models to the Estimation of School and Teacher Effects.* New York: AERA.

Webster, W.J. (In press, 1997a). "A Comparison of the Results Produced by Selected Regression and Hierarchical Linear Models in the Estimation of School and Teacher Effect." *Multiple Linear Regression Viewpoints.* Chicago: AERA.

Webster, W.J., and Olson, G.H. (1988). A Quantitative Procedure for the Identification of Effective Schools. *Journal of Experimental Education*, 56, 213-219.

Figure 5. Formative and Summative Indicators Available to Schools

Indicator	SIP Goals(s) [1]	Targeted Audience [2]	Purpose	Date Available
1. Texas Assessment of Academic Skills (TAAS) [percent passing]. The TAAS is a statewide criterion-referenced test administered at grades 3-8 and 10 in reading and at grades 4, 8, and 10 in writing. Spanish versions are available at grades 3 and 4. Science and social studies are tested at grade 8.	1, 2, 3, 4			
• disaggregated by demographic variables within school and district (demographic variables for reporting purposes are ethnicity, economic status, and English proficiency).		4,5,6,7,8,9	Accountability	Summer
• skills analysis				
√ by student		1,2,3	Decision-making	Spring,
√ by class		3		Fall
√ by teacher		3,4		
√ by school		4,5,6,7		
√ by district		6,7,8,9		
• reconstituted skills analysis			Decision-making	Fall
√ by student		1,2,3		
√ by class		3		
√ by teacher		3,4		
√ by school		4,5,6,7		
• teacher and school goals based on best practice		3,4,5,6	Decision-making	Fall
• school improvement/effectiveness indices		3,4,5,6,7,8, 9	Accountability	Summer
• teacher improvement/effectiveness indices		3,4	Accountability/ Decision-making	Fall

Figure 5. Formative and Summative Indicators Available to Schools

Indicator	SIP Goals(s) [1]	Targeted Audience [2]	Purpose	Date Available
2. *Iowa Tests of Basic Skills (ITBS/) Tests of Achievement and Proficiency (TAP)/ Spanish Assessment of Basic Education (SABE)* (NCE's and percentiles). These norm-referenced tests are administered at grades K-9 in reading and mathematics. Scores are reported relative to national norms.	1, 2			
• disaggregated by demographic variables within school and district		4,5,6,7,8,9	Accountability	Summer
• skills analysis (not available with survey form)			Decision-making	Spring
√ by student		1,2,3	(NCE's and	
√ by class		3	percentile also	
√ by teacher		3,4	reported at each	
√ by school		4,5,6,7	level that corres-	
√ by district		5,6,7,8,9	pond to the skills analysis.	
• reconstituted skills analysis			Decision-making	Fall
√ by student		1,2,3		
√ by class		3		
√ by teacher		3,4		
√ by school		4,5,6,7		
• teacher and school goals based on best practice		3,4,5,6	Decision-making	Fall
• school improvement/effectiveness indices		3,4,5,6,7,8,9	Accountability	Summer
• teacher improvement/effectiveness indices		3,4	Accountability/ Decision-making	Fall

Figure 5. Formative and Summative Indicators Available to Schools

Indicator	SIP Goals(s) [1]	Targeted Audience [2]	Purpose	Date Available
3. *Assessments of Course Performance (ACPs)* (scale scores). The ACP's consist of 72 standardized final exams in 72 high school courses, grades 9-12.	**1, 2,3,4,6**			
• scale scores by school		**3,4,5,6,7,8,9**	**Accountability**	**Fall, Spring**
• skills analysis			**Decision-making**	**Fall, Spring**
√ by student		**1,2,3**		
√ by class		**3**		
√ by teacher		**3,4**		
√ by school		**4,5,6**		
• teacher and school goals based on best practice		**3,4,5,6**	**Decision-making**	**Fall**
• school improvement/effectiveness indices		**3,4,5,6,7,8,9**	**Accountability**	**Summer**
• teacher improvement/effectiveness indices		**3,4**	**Accountability/ Decision-making**	**Fall**
4. Program Evaluation Reports. At least once every three years. These reports provide much of the information for school planning in that principals may pick and choose programs based on data. Example of program evaluation reports include evaluations of:	**1,2,3,4,5, 6,7,8,9, 10,11,12**	**3,4,5,6,7,8,9**	**Accountability/ Decision-making**	Summer, Fall
• **longitudinal achievement trends**				
• **accreditation status**				
• **TEA *AEIS* criteria**				
• **measurement profiles (*ITBS/TAP/SABE/TAAS*)**				
• **effective schools**				
• **graduate follow-up**				
• **school-centered education**				
• **learning centers**				
• **bilingual/ESL programs**				
• **reading improvement programs**				
• ***TAAS***				

Figure 5. Formative and Summative Indicators Available to Schools

Indicator	SIP Goals(s) [1]	Targeted Audience [2]	Purpose	Date Available
• magnet schools • gifted and talented programs • special math and science programs • ACPs • year-round education • the District Improvement Plan • dropout • services to at-risk student • teacher training • administrator training • Chapter I reading programs • parental involvement • migrant education • Chapter 2 block grants • programmatic remedies for low achieving students • drug usage • vouchers and charter schools • Even Start • community outreach • special education transition programs • *ACT/PSAT/SAT* results • enrollment trends and projections • head start transition				
5. Portfolios of Student Work (Development of portfolios currently being piloted in Chapter 1.)	1,2,3,4	1,2,3,4,5	Decision-making Used to supplement accountability information.	Continuous
6. Performance Tests (Development of Performance Tests being planned through the National Science Foundation Grant)	1,2,3,4	1,2,3,4,5	Decision-making. Used to supplement accountability information	Continuous
7. Teacher-made Tests	1,2,3,4	1,2,3,4,5	Decision-making. Used to supplement accountability information.	Continuous

Figure 5. Formative and Summative Indicators Available to Schools

Indicator	SIP Goals(s) [1]	Targeted Audience [2]	Purpose	Date Available
8. Systemwide Teacher Survey. This survey assesses teacher satisfaction with teaching, ranking of importance of educational goals, and perception of degree of seriousness of schoolwide issues. It is administered to all teachers.	12	3,4,5,6,7,8	Decision-making	Winter
9. Parental Involvement Log. A log, maintained at the school level, of parental involvement in school activities. Data are summarized at end of school year.	5	3,4,5,6,7,8	Decision-making	Continuous
10. Systemwide Volunteer Log. A log, maintained at the volunteer office, of volunteer hours and activities in each school. Data are summarized at the end of school year.	5	3,4,5,6,7,8	Decision-making	Continuous
11. Systemwide Parent Survey. This survey assesses parental school expectaions, perception of school climate, needs relative to the school, and involvement/ participation in school activities.	5,12	3,4,5	Decision-making	Winter
12. Chief Executive Officer (CEO) Management Report. This report, produced at the end of each six weeks, provides the principal, as CEO, with interim information on a wide variety of variables including: • financial allocations and expenditures • ACP results • Fall *TAAS* results • student promotion and retention rates • student enrollment in advanced courses, diploma plans, and honors programs • dropout trends • college entrance test participation • teacher attendance • student attendance • teacher vacancies • teacher grade distributions • student mobility	1,2,3,4,6,7, 8,10,11	4,5	Decision-making	Each 6 weeks

Figure 5. Formative and Summative Indicators Available to Schools

	Indicator	SIP Goals(s) [1]	Targeted Audience [2]	Purpose	Date Available
13.	**Teacher Climate Survey.** Available on request of the principal.	11,12	4,5	Decision-making	On request of principal.
14.	**Student Climate Survey, grades 4-12.** Available on request of the principal.	10,12	4,5	Decision-making	On request of principal.
15.	**Systemwide Student Survey, grades 4-12.** Systemwide survey assessing student satisfaction with learning, academic self-concept, family emphasis on education, cohesion. Administered once every three years.	10,12	3,4,5,6,7,8	Decision-making	Winter
16.	**Systemwide Principal Survey.** Assesses principal perceptions of effectiveness of training, effectiveness of central office support departments, school-wide issues, and decentralization.	12	4,5,6,7,8	Decision-making	Winter
17.	**School Improvement Effectiveness Indices.** • disaggregated by outcome and background variables (interactions included)	1,2,3,4,6,7, 8,9,10	3,4,5	Decision-making	Fall
	• by outcome variable and aggregate of all variables.		4,5,6,7,8,9	Accountability	Summer
18.	**Teacher Improvement Indices**				
	• by outcome variable and class		3	Decision-making	Fall
	• by teacher		4	Decision-making	Fall
19.	**SIP Goals and Goal Attainment**	1,2,3,4,5, 6, 7, 8, 9, 10, 11, 12	1,2,3,4,5,6, 7, 8, 9	Decision-making/ Accountability	Fall/Spring

1 SIP goals include 1) reading/language arts, 2) mathematics, 3) social studies, 4) science, 5) parental and community involvement, 6) promotion and course passing rates, 7) enrollment in advanced courses, etc., 8) dropout, 9) college entrance test performance, 10) student attendance, 11) teacher attendance, 12) climate and safety.

2 Targeted audiences include 1) student, 2) parent, 3) teacher, 4) principal, 5) School Community Council, 6) Central Office Line Officers, 7) Central office staff, 8) superintendent, 9) Board of Education and public. Any information that goes to the Board of Education and the public can also go to the Region, State, or National levels.

Figure 6. Description of Variables and Methodology Used In The School Effectiveness Indices

OUTCOME	GRADES	METHODOLOGY	POSSIBLE PREDICTORS	GRADES	SCHOOL LEVEL FAIRNESS VARIABLES
1. *Iowa Tests of Basic Skills*, year n, Reading and Mathematics (raw score)	1-8	HLM on residuals (student level)[1]	*Iowa Tests of Basic Skills*, year n-l, Reading and Mathematics	K-7	School mobility, school overcrowdedness, school level average family income, school level average family education level, school level average poverty index, school level percent students on free/reduced lunch, school level percent limited English proficient students, school level percent Black, Hispanic, and minority students, school level percent instructional days lost to medical disability leave and unfilled vacancies.
			Texas Assessment of Academic Skills, year n-1, Reading and Mathematics	3-7	
2. *Tests of Achievement and Proficiency*, year n, Reading and Mathematics (raw score)	9	HLM on residuals (student level)	*Iowa Tests of Basic Skills*, year n-1, Reading and Mathematics	8	Same as #1.
			Texas Assessment of Academic Skills, year n-1, Reading and Mathematics	8	
			Tests of Achievement and Proficiency, year m-1, Reading and Mathematics score (for students with insufficient credits to be in tenth grade)	9	

[1] HLM is Hierarchical Linear Modeling. HLM is a multi-level regression analysis technique which permits both student-and school-level data to be regressed against outcome variables simultaneously. See references Bryk & Raudenbush (1992) and Bock (1989). Whenever HLM is run on residuals, those residuals are obtained from student level OLS regression equations that are designed to control for the effects of gender, ethnicity, limited English proficiency, and socioeconomic status as well as appropriate interactions.

Figure 6. Description of Variables and Methodology Used In The School Effectiveness Indices

OUTCOME	GRADES	METHODOLOGY	POSSIBLE PREDICTORS	GRADES	SCHOOL LEVEL FAIRNESS VARIABLES
3. Promotion Rate, year n (percent promoted)	1-6, 7-8	Multiple regression (school level)[2]	Promotion rate in years n-1 and n-2.	1-6, 7-8	None
4. Student Attendance, year n (days attended)	1-12	HLM on residuals (student level)	Student Attendance, year n-1	K-11	Same as #1.
5. *Texas Assessment of Academic Skills*, year n, Reading and Mathematics (raw score)	3-8, 10	HLM on residuals (student level)	Texas Assessment of Academic Skills, year n-1, Reading and Mathematics	3-7	Same as #1.
			Iowa Tests of Basic Skills, year n-1, Reading and Mathematics	2-7	
			Tests of Achievement and Proficiency, year n-1, Reading and Mathematics	9	
6. *Texas Assessment of Academic Skills* Spanish, year n, Reading and Mathematics (raw score)	3,4,5,6	HLM on residuals (student level)	*Woodcock-Muñoz Language Survey, Broad Ability Score*, year n-1	2,3,4,5	Same as #1

[2] Multiple regression is a technique for predicting outcome variables based on related input variables. See references Draper and Smith (1968) and Aiken and West (1991).

Figure 6. Description of Variables and Methodology Used In The School Effectiveness Indices

OUTCOME	GRADES	METHODOLOGY	POSSIBLE PREDICTORS	GRADES	SCHOOL LEVEL FAIRNESS VARIABLES
7. *Texas Assessment of Academic Skills*, year n, Writing (raw score) (Spanish writing at grade 4)	4, 8, 10	HLM on residuals (student level)	*Texas Assessment of Academic Skills*, year n-1, Reading and Mathematics	3,7,9	Same as #1.
			Iowa Tests of Basic Skills, year n-1, Reading and Mathematics	3,7	
			Tests of Achievement and Proficiency, Reading and Mathematics	9	
			Woodcock-Muñoz Language Survey, Broad Ability Score, year n-1	3,7,9	
8. *Texas Assessment of Academic Skills*, year n, Science and Social Studies (raw score)	8	HLM on residuals (student level)	*Texas Assessment of Academic Skills*, year n-1, Reading and Mathematics	7	Same as #1.
			Iowa Tests of Basic Skills, year n-1, Reading and Mathematics	7	
9. *Spanish Assessment of Basic Education*, reading and mathematics, year n (raw score)	1-6	HLM on residuals (student level)	*Spanish Assessment of Basic Education*, year n-1, Reading and Mathematics	1-5	Same as #1.
			Woodcock-Muñoz Language Survey, Broad Ability Score, year n-1	K-5	

Figure 6. Description of Variables and Methodology Used In The School Effectiveness Indices

OUTCOME	GRADES	METHODOLOGY	POSSIBLE PREDICTORS	GRADES	SCHOOL LEVEL FAIRNESS VARIABLES
10. Appropriate ESOL ACP, year n (There are five ESOL ACPS, ESOL 1-3, Reading, and ESOL 1-2, Listening). Students should take the ESOL ACP that is appropriate for the course that they are enrolled in. These tests will be administered in the Spring.	7-9	HLM on residuals (student level)	*Woodcock-Muñoz Language Survey, Broad Ability Score*, year n-1	6-8	Same as #1
11. *Assessments of Course Performance*, year n, Language Arts (including ESOL, grades 10-12, first semester, grade 9, and first and second semester, grades 10-12), Mathematics, Social Studies, Science, Reading, World Language, (72 courses). Honors courses are considered separately. (raw score)	9-12	HLM on residuals (student level)	*Assessments of Course Performance*, year n-1, 72 courses, best predictors	7-11	Same as #1.
			Tests of Achievement and Proficiency, year n-1, Reading and Mathematics	9	
			Iowa Tests of Basic Skills, year n-1, Reading and Mathematics	7-8	
			Texas Assessment of Academic Skills, year n-1, Reading and Mathematics	7, 8, 10	
			Woodcock-Muñoz Language Survey, Broad Ability Score, year n-1	8-11	

Figure 6. Description of Variables and Methodology Used In The School Effectiveness Indices

OUTCOME	GRADES	METHODOLOGY	POSSIBLE PREDICTORS	GRADES	SCHOOL LEVEL FAIRNESS VARIABLES
12. *Woodcock-Muñoz LanguageSurvey, Broad Ability Score*, year n, (raw score)	1-6	HLM on residuals (student level)	*Woodcock-Muñoz Language Survey, Broad Ability Score*, year n-1	K-11	Same as #1.
13. Graduation Rate, year n (percent graduated)	9-12	Multiple regression (school level)	Graduation Rate, years n-1 and n-2	9-12	None
14. *Scholastic Aptitude Tests* and *American College Test*, Percent Tested, year n (percent tested)	9-12 (statistic is based on percentage of twelfth graders who have ever taken the test)	Multiple regression (school level)	Percent Tested, *Scholastic Aptitude Test* and *American College Test*, years n-1 and n-2	9-12	None
15. *Scholastic Aptitude* Test and *American College Test,*, Verbal and Quantitative, year n (raw score)	9-12 (student's best score)	HLM on residuals (student level)	Various *Assessments of Course Performance*, year n-1, Language Arts, Mathematics, Social Studies, Science, Reading, World Language (72 courses)	8-11	Same as #1.
16. Dropout Rate, year n-1 (percent dropout)	7-12	Multiple Regression (school level)	Dropout Rate in years n-2 and n-3.	7-12	None

Figure 6. Description of Variables and Methodology Used In The School Effectiveness Indices

OUTCOME	GRADES	METHODOLOGY	POSSIBLE PREDICTORS	GRADES	SCHOOL LEVEL FAIRNESS VARIABLES
17. Student enrollment in Prehonors and Honors Courses, year n (percent in accelerated courses)	7-12	Multiple regression (school level)	Student enrollment in Prehonors and Honors Courses in years n-1 and n-2.	7-12	None
18. Student enrollment in Advanced Placement Courses, year n (percent in advanced diploma plans)	11-12	Multiple regression (school level)	Student enrollment in Advanced Placement Courses in years n-1 and n-2.	11-12	None
19. Percent Tested, *Preliminary Scholastic Aptitude Test,* year n (percent tested)	9-12	Multiple regression (school level)	Percent tested, *Preliminary Scholastic Aptitude Test,* years n-1 and n-2.	9-12	None
20. *Preliminary Scholastic Aptitude Test*, year n, Verbal and Quantitative (scale score)	9-12 (student's best score)	HLM on residuals (student level)	Various *Assessments of Course Performance*, year n-1, Language Arts, Mathematics, Social Studies, Science, Reading, World Language (72 courses)	8-11	Same as #1.

Figure 6. Description of Variables and Methodology Used In The School Effectiveness Indices

OUTCOME	GRADES	METHODOLOGY	POSSIBLE PREDICTORS	GRADES	SCHOOL LEVEL FAIRNESS VARIABLES
21. Percent of students passing the Advanced Placement Examinations, year n (denominator is number of students enrolled in Advanced Placement Courses)	11-12	Multiple Regression (school level)	Percent of students passing the Advanced Placement Examinations, year n-1 and n-2.	11-12	None

CHAPTER 10

MONITORING AND EVALUATION OF EDUCATIONAL REFORM INITIATIVES IN THE STATE OF PARANÁ, BRAZIL

María Teresa de la Fuente, with Heloisa Luck and Corinna Ramos

One of Brazil's most advanced processes of educational reform was instituted by the state of Paraná. This case illustrates how evaluation activities have been linked to the implementation of new policies in an effort to develop innovations that will help schools consolidate their administrative systems and aid the state in reorienting its decisions and investments.

INTRODUCTION

Brazil has committed itself to carrying out a substantial reform of its educational system. This paper presents the context of the reform and describes initiatives for change being implemented in the country at the state level. To illustrate these initiatives, the paper highlights efforts currently being instituted in the state of Paraná, with special emphasis on innovations related to their monitoring and evaluation. Finally, this paper provides a discussion of the lessons emerging from the evaluation of educational reform in Paraná.

THE CONTEXT OF EDUCATIONAL REFORM IN BRAZIL

Since the 1980s, Brazil has been the focal point of fierce debates over the need for national educational reform. These debates, which are characterized by a focus on the democratization of education, have placed in motion a series of actions which, driven by various sectors of the country, are defining reform on all levels of the educational system. Initial actions made education available for virtually everyone: coverage was achieved for the large majority of the school-age population. This universalization occurred gradually, becoming a drawn-out process of transition that has posed challenges and problems for the country's educational systems, including:

- a scarcity of qualified teachers, leading to the hiring of teachers not always able to perform their assigned duties;

- management problems growing out of the need to expand services in order to satisfy increased demand; and
- the depletion of physical, financial, and human resources, creating deficiencies that render providing a high-quality education virtually impossible.

These problems are manifested in high rates of academic failure, which has in turn led to a disparity in the age-per-grade ratio in the system in general; a low rate of scholastic achievement, as reflected in substandard scores on standardized national tests; and a high dropout rate (72 percent of students fail to complete first grade) (Luck, 1996).

In Brazil's centralized educational system, all educational policy is dictated by the Ministry of Education (MOE), whose function is the "coordination of educational policy, by articulating the various levels and systems involved and performing a normative, redistributive, and supplementary function with regard to all other educational providers" (Article 8, para. 1, Law 9394/96).

At present, the MOE carries out its functions through three separate entities: the National Council on Education, the National Council of Secretaries of Education (CONSED), and the National Union of Municipal Directors of Education (UNDIME). These entities were created in order to make educational reform more dynamic and ensure the integrated development of Brazil's educational system.

The National Council on Education was created by Law 9193 on November 24, 1995. It is invested with normative and supervisory functions at several levels of the system and represents organized segments of society through its Chambers of Education. The National Council on Education replaces the former Federal Council on Education.

The National Council of Secretaries of Education (CONSED), a nonprofit organization chartered in accordance with private law, was established in September 1986. CONSED consists of the twenty-seven Secretaries of Education operating on the state and Federal District level. Its purposes are "to act as a permanent entity for coordinating and articulating the shared interests of the country's Secretariats of Education; to participate in the development and implementation of national education policies; to encourage and promote the quantitative and qualitative development of the country's public education system; and to establish positive interaction with all segments of political and civil society, with a view toward creating fair and equitable social relationships in a context of democratic management" (CONSED bylaws, 1995).

With a function similar to that of CONSED, but at the municipal level, the Union of Municipal Directors of Education (UNDIME) comprises all directors of education in Brazil's municipalities and is organized in accordance with the country's five geographic regions: Northeast, Northwest, Southeast, South, and West-Central. The primary purpose of UNDIME is to improve the quality of education in general and, as dictated by its legal mandate, of first-grade education in particular. The strategy adopted by UNDIME calls for integration among municipalities in order to build mutual strength in the search for results.

The joint efforts of these entities have led to an increased understanding of the different types of intervention needed to change the current status of education in Brazil. The priorities for change are outlined in Article 214 of the Federal Constitution of 1988, which provides for mandatory basic education throughout the country.

Guidelines for educational reform are applied to the Brazilian educational system in its entirety, which operates at three distinct levels: federal, state, and municipal. In accordance with existing legislation, these levels are interconnected with regard to the discharge of complementary and converging functions. The MOE dictates policy and coordinates activities at the national level for the development of basic education. It also conducts activities aimed at improving higher- and professional-level education. Responsibility for providing basic education falls to the states and municipalities, with the states responsible overall for coordinating this process. This coordination effort constitutes, in and of itself, a significant innovation in Brazilian education, since historically educational processes have been characterized by marked differences, a phenomenon that also affects the resources available at the various educational levels and geographic regions of the country.

Three legal documents currently define educational reform in Brazil, to wit, the Ten-Year Plan of Education for All (1993), the new Law of Guidelines and Bases for National Education (1996), and Law 9424 (1996). These documents resulted from the widespread mobilization of the various segments of society and constitute the most important guidelines for the implementation of national educational reform.

The first of these documents, the Ten-Year Plan of Education for All (1993-2003), drafted with the participation of most of Brazil's states and municipalities, enables "each federative unit and each municipality, based on the guidelines and goals set forth in the Plan, to define its own commitment and establish its own goals" (Cunha, undated). The Plan establishes eleven strategic lines of action, "taking into account the need to focus energies, media, and resources on improving education by encouraging completion of the full basic education scheme in order to eliminate illiteracy and the underschooling of youth and adults" (MOE, 1993:45). These lines of action, which have been approved at the national level, are listed below:

1. Establishment of basic standards for the public network
2. Establishment of minimum curricular content as determined by the Constitution
3. Professionalization and acknowledgment of the teaching profession
4. Development of new standards for educational management
5. Encouragement of innovation
6. Elimination of educational inequalities
7. Improvement in the degrees of access to schooling and length of time enrolled
8. Systematization of continuing education for youth and adults
9. Production and dissemination of knowledge and information in the field of education
10. Institutionalization of state and municipal plans
11. Professionalization of education administration

The second document, the Law of Guidelines and Bases for National Education (LGB), dated December 20, 1996, and identified as Law 9394/96, spells out norms governing the organization and operation of schooling in Brazil, at all levels and for all teaching modalities. It defines the various educational levels as follows: basic education (infant, primary and intermediate education) and higher education. It also defines the responsibilities of the various administrative levels (national, state, federal district and municipalities, school, and teacher), provides guidelines as to the professional training that must be incorporated into the various types of education, including school-based education; defines education for youth and adults as well as special education; and dictates the use of financial resources for education (López, 1977:41).

The third document is an outgrowth of the 14th Amendment to the Constitution that calls for the creation of a fund in each state and the Federal District. Approved on December 24, 1996, and identified as Law 9424, it regulates the funds to be allotted for education in the states and municipalities. This law significantly affects the scope of the objectives of the reform process, as it deals with the redistribution and transfer of financial resources for the maintenance and development of fundamental education (eight years of compulsory schooling), as well as for the evaluation of the teaching profession at this educational level, thus making it possible to resolve the financial imbalances among towns.

STATE INITIATIVES THAT RESPOND TO THE GUIDELINES FOR EDUCATIONAL REFORM

At present, the Brazilian educational system is faced with the urgent need to increase recorded levels of scholastic achievement. The problem of low academic performance affects a large number of Latin American countries. Brazil is among those with the lowest levels of achievement, together with Haiti, the Dominican Republic, Guatemala, El Salvador, and Honduras (Wolf, Schiefelbein and Valenzuela, 1994). Since "state systems of education are responsible for close to two thirds of the primary and intermediate school enrollment in Brazil, ... the Secretariats of Education play a significant role in the improvement of public education" (CONSED, 1996:64). These improvements are being achieved through the implementation of innovative projects in the state systems. "Innovative experiences are defined as those that introduce some type of change into a given academic culture and/or practice, through an intentional intervention or for a previously defined purpose. The innovation must be deliberate, planned, and carried out in a logical sequence in response to a previously defined purpose" (Pacheco Mendes, in Leonardos et al., 1994:10).

These innovations are classified in accordance with the following focal areas: increased coordination between states and municipalities; decentralization and increased autonomy for schools; training and assessment of educational professionals; improvement in the quality of education and school management; introduction of strategies for dealing with a variety of clienteles; and incorporation of educational technologies.

Table 3 (at end of this chapter) presents the innovative projects and activities being carried out in each of Brazil's 26 states. The table identifies "programs ranging from the use of school banks (*cajas escolares*) as instruments for strengthening school autonomy to the introduction of radical reforms in the organization of teaching" (CONSED, 1996).

The states are also demonstrating an increased interest in introducing monitoring and evaluation practices in an effort to identify the effects produced by such innovations. It is important to point out that this is a recent practice in Brazil, having been formally initiated in 1986 with the implementation of the National Assessment of Academic Achievement. Since 1990 the evaluation system has included "conceptual and academic factors impacting the quality of Basic Education provided to the country's various population segments and subsegments" (MED/SEDIAE/DAEB, 1996:20).

NATIONAL ASSESSMENT OF ACADEMIC ACHIEVEMENT

The National System for the Assessment of Basic Education (SAEB) assesses the levels of scholastic achievement attained by students and has as its objective obtaining information on the status of education in Brazil. Initial attempts at assessing academic performance began in the 1980s as an outgrowth of requirements imposed by international organizations, whose prerequisites for disbursing loans stipulated the implementation of such an evaluation structure. In 1990, the SAEB was administered to students from the first, third, fifth and seventh grades throughout the country. The results revealed that scholastic achievement was extremely low. Measurements of student mastery ranged from 30 to 56 percent, varying by region and by the area of the curriculum being assessed. In some cases, "less than one student per thousand successfully mastered the minimum content of the classes in which he or she was enrolled" (Nova Escola, 5/97).

The SAEB was administered again in 1993 and 1995. These results corroborated the low academic performance of the system in general and of the higher-level grades (fifth and seventh) in particular. They also showed that higher-than-average scores occur primarily in the South, Southeast, and West-Central geographic regions, while lower-than-average scores are concentrated in the country's North and Northeast regions.

PARANÁ STATE: EVALUATION AND MONITORING OF EDUCATIONAL REFORM ACTIVITIES

Paraná state is located in the southern region of the country and has an economy that is rapidly evolving from agricultural to industrial. Its population of 8.4 million comprises a variety of ethnic groups. According to data for November 1996, the state public education system has 1,274,767 students at the basic education level distributed among 2,046 school units. In the municipalities, students number 545,752 distributed among 6,824 schools, of which 1,454 are urban and 5,370 rural.

At present, the state public network consists of 44,521 employees responsible for satisfying the demand for basic education at all levels and in all modes. Of this total, 38,720 are classroom teachers, while 5,801 are educational professionals providing administrative and teaching support.

Compared to other states, Paraná is well-positioned with respect to the scholastic performance of its students as documented by the SAEB tests administered at the fourth and eighth grade levels of fundamental education and at the third level of secondary education. In 1996, Paraná ranked second in the country in Portuguese language (reading), with a score of 52.4 percent for the fourth grade of fundamental education, three percentage points above the national average. In mathematics for the same grade, it ranked third nationally, with a score of 31.2 percent, or 1.7 percentage points above the national average. With regard to academic achievement in the eighth grade of basic education, Paraná received a score of 67.4 percent for Portuguese language (reading) and ranked sixth as compared to other states, at a level 1.5 percentage points above the national average. It also scored sixth in mathematics, with an academic performance of 36.2 percent, which is 0.4 percent above the national average.

Also according to data from the SAEB, the evaluation of the third grade of secondary education in Portuguese language for 1996 shows a performance of 66.7 percent, earning it a seventh-place ranking. In mathematics for the same grade, Paraná achieved a scholastic performance of 36.7 percent and a ranking of sixth.

Even though the results obtained by Paraná state are not the lowest in the country, the Paraná Secretariat of Education (SEED-PR) is introducing innovations designed to improve the quality of basic education. These innovations are grouped in accordance with three distinct norms as described in the Action Plan of the Secretariat of Education: 1995-1998, as follows: "*the student* remains successfully enrolled in school, experiencing new and significant educational opportunities; *good teachers* develop their professional, personal, and cultural skills on a systematic and ongoing basis; and *the community* participates effectively in decision mak ing in conjunction with the system for achieving educational objectives" (SEED-PR, 1995).

These norms require the identification of solutions that will increase the levels of scholastic achievement by students; institute management practices in both schools and the community; and establish a system of ongoing practical and in-service training for teachers and educational professionals. In support of these norms, the SEED-PR has already initiated a process for implementing monitoring and evaluation activities that are at present in either particular stages of development or in their second year of operation.

MONITORING AND EVALUATION OF ACADEMIC ACHIEVEMENT IN PARANÁ

In Paraná State, assessment of academic achievement differs from the method espoused by the MOE. All state schools, and by default all municipal and private schools as well, are covered. The principle of universality, as opposed to the principle of sampling, is favored in order to incorporate all schools into the practice of monitoring and evaluating the quality of education. The principle of universality also allows each teacher, school, and community in the state to be included in the aggregate data for the state system, thus promoting their own self-evaluation.

The assessment of Paraná's educational system forms a part of the institutional development component of the Paraná Quality of Public Education Project (PQE), an initiative co-financed by the World Bank and the Government of Paraná State. It includes the evaluation of academic achievement in primary and intermediate education. The program has the following stated objectives:

- to provide the State Secretariat of Education with up-to-date information on academic performance;
- to determine school performance in the areas of organization, management, and coordination with the community; and
- to promote the development and practice of monitoring and evaluating school and educational management so as to improve education within the state.

The ultimate goal of this program is to establish a flexible system of monitoring and evaluation so that users can draw directly on the data and information gathered and use it for their own self-evaluation and development. A related goal is to allow users to define appropriate educational activities for the local, regional, and central levels. The evaluation is conducted as a collective project requiring the effort—and the participation—of students, parents, directors, school technical-pedagogical teams, and representatives of the Regional Nuclei, all coordinated by the Ministry of Education.

Parents participate as testing monitors so that they can not only contribute to the evaluation but also—and above all—validate the assessment as an essential instrument for verifying student learning. Historically, evaluation has been considered an instrument of power wielded by teachers and used to determine which students are qualified to graduate to the next level and which students will be held back. Regional Nuclei, intermediate administration units between the Ministry and the schools who act through the members of their technical teams, function as regional coordinators in the process.

Grades and subjects evaluated

Grades and subjects are evaluated in accordance with the implementation program that is conducted and supervised on a universal basis (Table 1).

Table 1. Evaluation Implementation Program. Grades and Subjects Evaluated

	1995	1996	1997	1998
Grades	Fourth grade of fundamental education	Eighth grade of fundamental education	Fourth grade of fundamental education	Fourth and eighth grades of fundamental education and second grade of intermediate education
Subjects	Portuguese and mathematics	Portuguese, science, history/geography and, for the second grade of intermediate education, Portuguese language and mathematics	Portuguese, mathematics, science, history, and geography	Portuguese, science, history/geography, and, for the second grade of intermediate education, Portuguese language and mathematics

Simultaneous with testing academic achievement, questionnaires are being applied with regard to the pedagogical practices of teachers, the function of school management and organization, and the coordination with the community. These questionnaires are intended for management personnel, the technical-pedagogical team, and classroom teachers. They are designed to verify the relationship between differences in rates of academic achievement and the nature and systematization of daily classroom practices.

Development of tests

Tests are developed by teachers affiliated with the State Education Network who have experience in the grades being evaluated. These teachers are supported by representatives of the Departments of First- and Second-Grade Education of the SEED-PR, and act under the coordination of a consultant from an Institution of Higher Education (IES).

The test development process is intended to ensure that the tests represent the curriculum taught and promote teacher participation. In addition, the process is an attempt to ensure that teachers accept the tests and use the results.

Scoring of tests

Scoring of the tests is carried out in each school by a team of teachers and testers and by evaluation monitors, all specially appointed and trained for that purpose. The procedures involved in this phase are detailed in the *Test Scoring Manual*, and are the respon-

sibility of the school director, who also receives special training for this purpose. The option for scoring tests in the schools themselves was established in order to provide immediate information to the school, as well as to increase teacher participation in the monitoring and evaluation process.

Results and uses for improving academic achievement

The results of the evaluation of academic achievement obtained in 1995 and 1996 have identified critical areas similar to those revealed by the SAEB. Each school receives the results of each subject evaluated, thus enabling it to identify those aspects of learning characterized by critical failings and constraints. In this way, each school can determine the particular interventions it requires to correct any deficiencies. It is expected that future evaluations of the same grade will make it possible to supervise and evaluate, on a comparison basis, the effectiveness of these interventions. The result of these tests will also enable the SEED-PR training program to target its action programs to areas requiring strengthening.

MONITORING AND EVALUATION OF THE CONTINUOUS TRAINING OF EDUCATION PROFESSIONALS

Since 1995, the ongoing training of SEED-PR education professionals has been conducted along two lines of action: Continuous In-Service Training and Theoretical-Practical Improvement. These lines of action have been included in programming for the Universidad del Profesor, created as an "institution of support for education, with the goal of conceiving, developing, and implementing activities involving the training of teaching and support personnel" (State Government/SEED-PR, 1995, bylaws of the Universidad del Profesor, Art. 2) through associations with other public or private institutions. Created on October 20, 1995, and officially chartered on October 28 of the same year, it began operating in 1996 with financial support provided by the Paraná Quality of Public Education Project (PQE).

The Universidad del Profesor program includes in its training activities content and methodologies taken from both formal and nonformal education. The program is aimed at promoting personal improvement as well as the in-service training of education professionals involved in public and municipal education in Paraná state. Training is not mandatory. All professionals affiliated with the system are free to participate and do so in response to their degree of professional motivation and needs. The primary objectives of the Universidad del Profesor are as follows:

- to provide opportunities for teachers to develop their personal, professional, and cultural skills to prepare them to perform efficiently on both the individual and collective levels;
- to implement strategies of shared responsibility for managing the process of ongoing in-service training and theoretical-practical improvement of education professionals;
- to encourage, within the framework of a professional environment, the communication of innovative experiences arising out of daily occurrences in the school environ-

ment, in order to further disseminate and implement ideas that can provide solutions to the problems of education in Paraná state; and
- to train education professionals in specific areas of the curriculum as well as in the systematization of pedagogical and school management experiences.

The Universidad del Profesor training program consists of three training modes: seminars, courses, and in-school training. Two types of seminars are provided: an in-service and motivation seminar, also known as an "advanced education" seminar, and an in-service seminar on curricular content, school management, and support for curricular practice. The courses offered are also of two types: specialization courses and extension courses with technical-pedagogical counseling. The third mode, in-school training, involves four types of learning: study groups, tele-classrooms, pedagogical encounters, and continuous training modules.

Salient to the in-school training mode are *study groups* (in which the participants, from either one or a number of schools, decide on both the subject matter and frequency of the study sessions); *tele-classrooms* (conducted in the various tele-halls distributed throughout state schools and Regional Nuclei [NR] using long-distance training techniques, videos and, in the future, the Internet); *pedagogical encounters* (non-regular events convened to address specific subjects selected on the basis of the school's needs and requirements); and *continuous training modules* (conducted by the teachers and focused on teachers in the school and in specific curricular areas).

Monitoring and evaluation team

As part of the PQE Project, a Monitoring and Evaluation Team was created at the central level of the Ministry of Education whose principal function is to complement the Teacher Training Program while at the same time to initiate and implement monitoring and evaluation within the state school system. In operation since 1995, the team monitored and evaluated training during 1996 for the purpose of providing the SEED with qualitative and quantitative information on the training provided. The team seeks to answer three questions: Which activities are being implemented appropriately? How do participants evaluate activities? And, what is the relationship between participation in training activities and teacher behavior in the classroom?

The information obtained through these formative monitoring and evaluation activities was used to plan new activities. Participant opinions and evaluation helped redesign certain strategies. Reproduced below are comments taken from the newsletter (November 1996) published by the Monitoring and Evaluation Team:

> Almost all (92 percent) of the participants filling out evaluation questionnaires indicated an extremely high level of satisfaction with the *In-service and Motivation Seminars* (*Advanced Education*). During the visits to schools, an effort was made to determine whether the effects of this seminar translate into concrete actions that benefit students. It may be said that benefits occur primarily in the areas of personnel and

> human relations. With reference to the seminars in curricular areas, it was observed that "a large majority of the participants feel that the contents were well presented and consistent with the proposed curriculum (an average of 86 percent, with a level of 97 percent achieved in the area of mathematics)."

At present, all levels of the educational system participate in the process of monitoring and evaluation of training, through functions ranging from data collection and analysis to dissemination of results. The idea is to implement a decentralized and collective process of feedback that will make it possible to ensure the qualitative and quantitative monitoring of training.

MONITORING AND EVALUATION OF EDUCATION MANAGEMENT: IN SEARCH OF EXCELLENCE IN SCHOOLS

> An educational institution oriented toward the integral development of students, educators, and community is based on the coordination of school and life, practice and theory, knowledge and work. Its challenge is to progressively develop programs and activities that will respond to the interests and expectations of the community with regard to the ongoing improvement of its administrative, pedagogical, and didactical processes, the constant reflection on its practices, and its dynamic interaction with its environment. In this sense, the educational institution is the guiding force of its own development since it is able to reorient academic processes based on the participation and responsibility of all of its actors (Salazar Ramírez and Quintero Gutiérrez, 1993:24).

The quality of school management is a determining factor in the quality of teaching since it conditions the functioning of the school, not only in terms of the attainment of its educational objectives but also in terms of its coordination with the community. School management in search of excellence consists of a process linking together multiple factors which, either directly or indirectly, are representative of school life in all its dimensions. Given its complexity, this interconnection demands of school directors a combination of skills that they do not always possess, given that Paraná school directors are elected by the direct vote of educators and the external community (parents and students), based on appreciations of management skills that are not always well-founded.

In Paraná state, the directors are teachers who are elected to serve for two years and who may be re-elected for another two years. In the most recent elections, which took place in 1995, a requirement was introduced that stipulated that each candidate prepare an action plan and submit it to the school community. Even though the election may reflect progress in democratizing public schools, it does not necessarily guarantee effective management practices, since candidates for the post receive no specific preparation for discharging that function. To counter this situation, the SEED-PR developed

specific, complementary lines of action. One of these is the training of administrators through in-service workshops on school management. Another is the so-called Project of Excellence and the creation of the Award for Excellence in Schools.

The award for excellence in schools

This award, which is based on the Malcolm Baldrige Model (Education Pilot Criteria) of 1995, attempts to provide a series of incentives to the system as a whole, based on a set of fundamental values and concepts for achieving excellence in the school. The requirements include the development of thinking skills, the improvement of the quality of teaching, and the acceptance of an objective process of school self-evaluation. Based on these requirements, a reference chart containing eight performance criteria was established. These criteria serve as guidelines for management activities and at the same time make it possible to conduct both an internal and external assessment of the school.

The performance criteria are as follows: leadership; information and analysis; strategic and operational planning; human resource development and management; management of educational processes, operations, and associations; maintenance of the school environment; results of school performance; and focus on the student and satisfaction of internal and external communities.

In order to support all state schools in an initial implementation phase and in anticipation of its extension to municipal networks, a process designed to seek excellence in schools was established. This process was launched in 1995 through the issuance of an invitation to schools to voluntarily discuss their priorities and formulate their own "project of excellence" based on a questionnaire distributed to all schools. In this way, schools had an opportunity to define for themselves the path they would take in their search for the desired degree of quality. The goal was to transform each director's Action Plan—submitted during his or her candidacy—into a project of excellence, once it had been reviewed and accepted by the school community.

Following identification by the school of its own priorities, initiatives to address those priorities, and indicators of the successful implementation of the school's plans and projects, the director and the teachers fill out a self-evaluation form and, on a voluntary basis, register for the award.

The award process does not limit the number of schools eligible to participate. A commission of external evaluators visits registered schools and prepares an analytical document that is then delivered to each school. This provides information on the administrative-pedagogical management practice as reflected in the current condition of the schools. The training of directors takes place simultaneously with the training of representatives of the Parent and Teacher Associations (PTA). The stated performance criteria—with their corresponding scores (on a scale of 1,000 points)—are as follows:

leadership (80 points); information and analysis (60 points); strategic and operational planning (50 points); human resource development and management (120 points); educational process, operations and associations management (120 points); maintenance of the school environment (70 points): results of school performance (250 points); and focus on the student and satisfaction of internal and external communities (250 points).

The first phase of the process of school self-evaluation concluded in November 1996. A total of 1,659 schools (82 percent of the total) participated in the process. Data tabulation reveals widely divergent results, with scores ranging from 100 or 200 points (in 1.7 percent of all schools) to 900 points (in one school). The greatest concentration of points was observed at a level of about 500, indicating a typical self-evaluation score located at approximately the midpoint of the maximum total score (26 percent of all schools), followed by a cluster of schools in the 400 range (24 percent of all schools), and finally by a third group in the 600 range (19 percent of all schools).

As shown in Table 2, the data indicate that the criterion by which the schools judge themselves in the most favorable light is that referring to school environment. It is also interesting to note that it was ultimately performance criteria, involving the pedagogical dimension, that received the lowest scores.

USEFULNESS, FEASIBILITY AND SIGNIFICANCE OF MONITORING AND EVALUATION PROCEDURES IN PARANÁ

The need to evaluate scholastic achievement in all of its dimensions—as well as to measure the progress achieved by the system in general and the schools in particular as a result of the proposed changes—has made it imperative to establish monitoring and evaluation activities. Within the context of the Brazilian educational system, there has been no tradition of developing integrated monitoring and evaluation systems. It has only been in recent years—as a result of actions such as the approval of new laws affecting education, the establishment of the SAEB, and the new initiatives promoting the increased systematization of school management—that it has been possible to perceive the usefulness and feasibility of introducing monitoring and evaluation practices aimed at facilitating decision making on the basis of objective, up-to-date information.

At present, there are several types of proposals within the various states for supervising and evaluating innovative activities in state educational systems. Most of these proposals are currently in the development and adjustment stage and represent initial steps toward the creation of more complete and integrated systems. However, even though the results obtained are, in the best of cases, only partial, sporadic information on specific initiatives is already being provided and used to redirect activities during program implementation.

Table 2. Average Scores for School Self-Evaluation Based on Performance Criteria and Scores as a Percentage of Maximum Possible Scores

Classification/performance criteria	Average absolute scores	Maximum possible score	Scores as a percentage of maximum possible scores
1. Maintenance of school environment	49.29	70	70
2. Leadership	55.81	80	69
3. Human resource development and management	73.48	120	61
4. Management of educational and operational processes and associations	73.76	120	61
5. Strategic and operational planning	30.37	50	60
6. Information and analysis	35.33	60	58
7. Focus on students and satisfaction of internal and external communities	134.45	250	54
8. Results in terms of school performance	133.09	250	53
TOTAL	585.58	1,000	

Source: SEED-PR, 1996.

Efforts made within the context of the Paraná experience have led to the following achievements:

1. **Increased participation by the educational community in the processes of monitoring and evaluation.** Participation in the processes of monitoring and evaluation helps define the significance of these activities in daily practice. "From the strategic standpoint, one of the most effective means for ensuring institutional independence . . . is precisely the articulation of a participative mechanism to preclude the existence of a single evaluation authority. To the extent that all sectors of the educational community are alternately both subjects and agents of evaluation as well as participants in the programs being implemented, it is possible to ensure true independence through the balancing of power" (Tiana, 1996). Active participation in evaluation processes can be observed in Paraná in the following activities: teacher-designed tests for evaluating performance in terms of specific areas of the curriculum to be evaluated, participation of parents and teachers in the testing process, scoring of the tests within the school, and subsequent distribution of results to schools to enable the latter to decide which corrective measures to apply.

Active participation can also be found in the area of training, as all individuals involved evaluate training activities. All modes of training have been supervised and evaluated by the SEED-PR monitoring and evaluation team. The team used questionnaires and interviews of both trainees and trainers; made visits to schools to observe the results of training in pedagogical practice; and created focus groups that made it possible to obtain a variety of perspectives with regard to the training.

Finally, active participation can be observed in the award of excellence project, beginning the moment projects are selected, implemented, and self-evaluated with the involvement of directors, teachers, and community. The results obtained to date in the area of participation are promising. Although participation is voluntary in both training activities and the award of excellence project, high levels of participation were achieved in 1996 (totaling 71 and 82 percent respectively).

2. Increased dissemination of significant information. "The production and dissemination of significant information on the performance of the educational system is one way to promote the participation in, and commitment to, education by the series of actors involved in it. In addition, it promotes an increase in the learning capacity of all levels of the system itself: the ability to know what is happening, to innovate and develop alternative strategies, and to systematically evaluate their results" (Toranzos, 1996). In the three above-mentioned initiatives, this dissemination has proven itself to be useful, feasible, and significant for all those involved. The results of the school performance evaluations that are distributed to each school make it possible for each to rate itself relative to the entire state and to propose its own corrective measures. The Monitoring and Evaluation Newsletter produced by the team provides information on training activities already implemented and in operation. With regard to the award for excellence in schools, it is expected that once the award has been granted, it will be possible to disseminate information about positive experiences and recognize the local efforts of all schools, in particular those with outstanding performances.

3. Use and integration of information. Within the context of the training program, it has been possible to use the information gathered, and to modify certain activities in response to trainee suggestions and needs. For example, the teachers deemed the practical workshops to be "the training mode that most contributed to their practice in the classroom," and in addition indicated that the "lectures and round table discussions were of little significance" (Newsletter, November 1996). With this information, the programming of training activities was reoriented to include more practical workshops. It has also been possible to use similar information to identify infrastructure problems that hamper rapid communication between the central, regional, and local levels and to devise temporary solutions for those problems. Future plans call for this information to be integrated into a system and for that system to generate a culture of monitoring and evaluation within the SEED-PR. Under the leadership of the SEED-PR's Department of Fundamental Education, research on integrating the monitoring and evaluation activities of the various programs of the Ministry has already begun.

4. The creation of the Universidad del Profesor. This has produced greater-than-expected benefits. Bringing together a large number of participants in one place has not only promoted monitoring and evaluation activities but has also made it possible to:

- standardize training to a large extent, as standards remain constant for all trainees. In the past, training was conducted by different providers in different areas of the state. This led to substantial variations in training, content, and quality.

- focus training on specific areas for each group. In this way, the training becomes both significant and dynamic while responding to the needs of the system.
- create opportunities for professional exchange among members of the teaching staff of the SEED-PR. "This makes it possible not only to provide in-service training but also to promote the exchange of experiences among teachers, which is most desirable since interaction and communication among small groups of teachers promotes in the latter a change of attitude vis-à-vis the introduction of new methodologies" (Colbert, 1990).
- facilitate training monitoring and evaluation activities, as it is possible to gather data on the shared experiences of a large number of participants. One additional benefit of the data gathering activity is that participants acquire practical experiences that sensitize them to monitoring and evaluation activities.

5. The creation of the Monitoring and Evaluation Team for the training component. This team has served as a link between the various levels of the system, communicating regularly with the local, regional, and central levels. It has conducted formative evaluations of the training program that have led to the introduction of improvements and adjustments to its activities and has provided regular training to regional representatives responsible for the monitoring of training activities conducted in the schools. During 1996, the team conducted these activities on a centralized basis from the state capital city of Curitiba, with regional representatives traveling to the capital city to receive training. In 1997, to provide a broader response to demand and in recognition of regional differences, Paraná state was divided into nine "poles" that serve as action centers. Based on this structure, the team traveled to the regions to meet the needs for technical assistance, training, and monitoring in the "poles." The results of these two strategies were compared at the end of 1997.

To implement an efficient system of monitoring and evaluation, the following conditions need to be created:

- a clear understanding of the concepts of monitoring and evaluation and of the objectives being pursued with such systems. Supervision or evaluation for their own sakes cannot be justified. These activities should form part of a larger plan aimed at improving the quality of education;
- a clear definition of the goals to be achieved and of the indicators to be used; in addition, qualitative and quantitative criteria must be established to make it possible to measure whether the goals were achieved;
- adequate provision of resources and incentives to motivate and facilitate achievement of the goals. This presupposes the design of flexible implementation strategies that will make it possible to address differences, introduce changes, and establish corrective measures, as necessary;
- an infrastructure that will permit the efficient gathering and analysis of information; and
- a plan for disseminating information throughout the system.

To establish such conditions requires time and resources that are not always available, at

least not in the appropriate amounts, and for this reason one of the principal challenges in implementing monitoring and evaluation activities is acquiring the physical and human resources to carry them out.

As is the case in other states, the criteria for evaluating the usefulness, quality, significance, and feasibility of monitoring and evaluation actions in Paraná are in various stages of development. Although specific methodologies are being used on a limited basis, current trends point to the development of a complete and integrated system in the future.

CONCLUSIONS RELEVANT TO OTHER CONTEXTS IN BRAZIL AND TO COUNTRIES FACING SIMILAR CHALLENGES

The initiatives proposed as part of the educational reform process currently underway in Brazil have created a legal, financial, and political framework that provides flexibility and support for establishing innovative measures aimed at improving the Brazilian educational system. The projects described in Table 3 reflect the consensus and awareness that exist today with regard to the need for change. They identify the areas in which such change is required. This convergence enables the entire system to benefit from each state's individual efforts. The government initiative has been, and will continue to be, a necessary condition for the emergence of common responses in pursuit of educational improvement in the country. The creation of a favorable climate for change is dependent, in great measure, on the leadership that governments are able to provide on all levels of the educational system.

The implementation of monitoring and evaluation systems presents, in and of itself, a significant challenge, first as a result of the lack of understanding with regard to the function of monitoring and evaluation in the educational environment in general, and second as a result of the lack of human resources and infrastructure for their implementation. As with all processes of change, prior to proceeding with the actual establishment of such systems, it is necessary to introduce a sensitization stage. The creation of a Monitoring and Evaluation Team within the organization has made it possible in Paraná to create conditions that sensitize and train its citizens in the use of such practices.

One of the greatest challenges involves securing the human and financial resources necessary to efficiently support a system of monitoring and evaluation. Efficiency is a necessary condition in such circumstances. It is for this reason advisable to begin with limited programs that can serve as pilot experiences and subsequently be scaled up when the necessary resources become available and new strategies have been adequately tested.

The integration of the various monitoring and evaluation activities is a long and difficult process, requiring changes to be made to established management processes as well as to the attitudes of individuals who produce and use the information. These changes are made gradually by means of adjustments and negotiations with regard to daily school affairs. Accordingly, they require time to take root and transform themselves into complete systems at the service of the entire educational organization.

REFERENCES

Brazil. Ley, Decretos, etc. 1995. Ley nº 9131/95: Altera los Dispositivos de la Ley nº 4024/1961. Lex: Coletânea de Legislação Federal. São Paulo: LEX, v. 59: 2042-2045, Oct./Dec.

Brazil. Ley, Decretos, etc. 1996. Ley nº 9394/96: Establece las Directrices y Bases de la Educación Nacional. Diario Oficial da Uniao (secão I), Brasilia, v.89, n. 248: 27833-27841, Decemer 23.

CONSED (Conselho Nacional de Secretários de Educação). 1996. *Relatório de Gestão 1995-1996.* Brasilia, Brazil: author.

CONSED (Conselho Nacional de Secretários de Educacão). 1995. *Estatuto y Regimento.* São Paulo, Brazil: author.

Cunha, C. Undated. "Do Plano Decenal: Compromiso e Alcance." In *MEC. O Plano Decenal Brasilia,* s.d.

Gobierno del Estado/SEED-PR. 1995. *Plano de Ação da Secretaria de Estado da Educação do Paraná 1995-1998.* Curitiba, Brazil: SEED-PR.

Leonardos, A.C., et al. 1994. *Estudo de Caso Aplicado às Inovações Educacionais: uma Metodologia.* Brasilia, Brazil: INEP.

Luck, H. 1996. *Community Involvement in the Development of Education: The Paraná Case.* Annual Conference of the Comparative and International Education Society, Williamsburg, VA: March 1996.

Nova Escola. 1997. *Avaliação Permite Cutucar Os Pontos Fracos Das Escolas.* Paraná, Brazil: Revista Nova Escola, Maio/97.

MEC. 1993. *Plano Decenal de Educação para Todos.* Brasilia, Brazil: MEC.

MEC/SEDIAE/DAEB. 1996. *Resultados do SAEB/95: Relatório Final.* Brasilia, Brazil: DAEB.

Salazar Ramírez, A., and Quintero Gutiérrez, R. 1993. *Una mirada cualitativa a la institución escolar.* Bogotá, Colombia: Ministerio de Educación Nacional.

Tiana, A. 1996. *La Evaluación de los Sistemas Educativos.* Madrid: Revista IberoAmericana de la Educación: 37-51.

Toranzos, L. 1996. *Evaluación y Calidad.* Madrid, Spain: Revista IberoAmericana de Educación: 63-78.

Wolf, L.; Schiefelbein, E.; and Valenzuela, J. 1994. *Improving the Quality of Primary Education in Latin America and the Caribbean.* Washington, DC: World Bank.

Table 3. Trends in Innovations in Brazilian States

STATE	PROJECT NAME	PURPOSE	YEARS IN IMPLEMENT-ATION
ACRE	*OPEN BOOK*	*Series of general measures implemented by the current administration:* Election of directors since 1981. Pioneer state for this initiative. School kits to combat truancy. Teacher training and bilingual/ intercultural education for indigenous populations.	Current
ALAGOAS	*INTER-INSTITUTIONAL ASSOCIATION*	*Integrated and shared school management* to provide service to abandoned children and their families. Development of associations among three levels (government, civil society, and private sector initiatives) with active community participation.	Since 1993
AMAPA	*SCHOOL BANKS*	*Decentralized school management* providing for the administrative, financial (*cajas escolares* or school banks), and pedagogical autonomy of the school. School banks are being established in 129 of the 467 state schools. They are designed to provide for the maintenance and conservation of school buildings and to provide school lunches and functional, didactic, and pedagogical materials.	Since 1995
AMAZONAS	*LIGHT OF KNOWLEDGE*	*School-boat to serve isolated communities of the Amazon.* The boat provides health services, schooling, teacher training, and community activities. One boat is in operation and eight are under construction.	Since 1996
BAHIA	*MEDIA REFORM OF INTERMEDIATE EDUCATION*	*Reform of intermediate education* by offering professional training courses in intermediate education.	Since 1996

STATE	PROJECT NAME	PURPOSE	YEARS IN IMPLEMENT-ATION
CEARA	*ALL FOR EDUCATION*	*Democratization of management* through the direct election of the directors of state public schools; increased financial autonomy of schools through participation in the School Development Support Fund (FADE); and administrative decentralization through the elimination of fourteen regional delegations. Centers directed by educators elected by the community.	Since 1996
FEDERAL DISTRICT	*BOLSA-ESCUELA*	*Promote promotion of admission and continued enrollment in public school with appropriate levels of academic achievement for students between the ages of 7 and 14 years* with precarious family and physical situations. The program received awards from UNICEF and the Getulio Vargas Foundation.	Since 1995
ESPIRITU SANTO	*DEBUREAUCRATIZATION AND DEMOCRATIZATION OF SCHOOLS*	*Democratic schooling through shared management.* Social integration and de-bureaucratization of educational information.	Since 1995
GOIAS	*TEACHER IMPROVEMENT*	Recognition of the courses of the teaching profession to *improve the qualifications of teaching personnel.* Implementation and systematic coordination are achieved through associations established between the superintendencies of primary and intermediate education, the schools of education of the University of Goias, the Department of Education of the Catholic University of Goias, and coordination with other projects.	Since 1991

STATE	PROJECT NAME	PURPOSE	YEARS IN IMPLEMENT-ATION
MARA-NHAO	*RECOVERY OF PUBLIC SCHOOLS*	Implements *innovative actions aimed at recovering public schools,* including accelerated studies for students with a difference in excess of two years in terms of age-for-grade.	Current
MATO GROSSO	*UNIFORM SYSTEM OF BASIC EDUCATION PROJECT—OUR SCHOOL*	With the creation of the Mato Grosso Educational Foundation, all *educational functions will be absorbed by that institution.* With advisory assistance from the Paulo Freire Institute, the organization is managed by a tripartite commission representing the states, municipalities, teachers/administrative personnel, and students. Each school creates its own strategic development plan in accordance with its own needs. The project evaluates the entire system.	Current
MATO GROSSO DEL SUR	*LITERACY TIME*	*Reduction of illiteracy* to a level of 5 percent in two years for illiterate individuals between the ages of 15 and 60. With technical advisory assistance from UNICEF and in association with dozens of organizations providing physical space and instruction.	Since 1996
MINAS GERAIS	*DEMOCRATIZATION OF SCHOOLS*	*Democratization of schools* through community election of directors; expansion of student body powers; teacher training; pioneer program for evaluating public schools.	Since 1991
PARA	*INTEGRATED SCHOOL-COMMUNITY PROJECT*	*In order to address the high rates of violence in state schools in Belem and transform schools* into community centers, schools offer activities and practical workshops in art, education, sports, and income generation, in addition to subjects of community interest. The results are promising and the project is being extended to other cities.	Since 1995

STATE	PROJECT NAME	PURPOSE	YEARS IN IMPLEMENTATION
PARAIBA	*PARAIBA CENTERS FOR SOLIDARY EDUCATION, OR CEPES*	*Created to recover the quality of education, the value of the teaching profession and teacher salaries, and educational efficiency,* the CEPES are located near schools in accordance with the number of participants.	Since 1996
PARANA	*NORMAL UNIVERSITY*	The largest program of *training for teachers of primary and intermediate education* in Brazil. Training modes include in-service training, training for teachers, and in-depth training in specific areas of the curriculum. Monitoring and evaluation of training already provided.	Since 1996
PERNAN-BUCO	*NETWORK TRAINING*	For the purpose of *ensuring consistency in training opportunities,* three types of training are offered: traditional training in association with universities and scholarships; continuing in-service training; and spontaneous and long-distance training. A counselor in the core school directs the discussion of the videos. The network evaluation provides resources for future programming.	Since 1995
PIAUI	*BY BUILDING, I LEARN*	Project to combat *the high degree of illiteracy* in the state (as much as 71 percent in Simões state). The project is implemented in association with the MOE, dioceses, prefectures, corporations, universities, and communities. The goal is to provide literacy training to 30,000 youths (over age 14) and adults, prior to 1998, in 70 percent of the state's inland municipalities.	Since 1996

STATE	PROJECT NAME	PURPOSE	YEARS IN IMPLEMENT-ATION
RIO DE JANEIRO	*REFORM OF INTER-MEDIATE EDUCATION*	Reform of intermediate education, in response to the need to provide general culturization and professional training to students graduating from the eighth grade of primary school. Referral centers offer courses in professional skills development.	
RIO GRANDE DO NORTE	*PRO-TEACHING PROJECT*	Pilot project in technical cooperation with France for the training of teachers based on the existence of a tutor-student relationship.	Since 1994
RIO GRANDE DO SUL	*SEDUCTIVE SCHOOL*	Schools incorporate art into the curriculum in order to provide opportunities for students to experience artistic manifestations. These manifestations are integrated into academic knowledge *in order to facilitate personal and school autonomy.*	Current
RONDONIA	*CLASSROOM SUPPORT*	Created to *facilitate individualized attention* in accordance with the specific needs of each student. The program provides pedagogical support to students showing low academic achievement in order to prevent failure. The program is being expanded in the capital city and to municipalities.	Since 1996
RORAIMA	*TEACHER TRAINING FOR INDIGENOUS DOCENTS*	*Teacher training* for lay indigenous individuals based on the conservation of their ethnic and social culture.	Since 1994
SANTA CATARINA	*IMPROVING THE QUALITY OF EDUCATION*	*Actions aimed at improving the quality of education,* including: the re-evaluation of educational proposals, teacher training, restructuring of intermediate education, and focus on the areas of citizenship and work.	Since 1989

STATE	PROJECT NAME	PURPOSE	YEARS IN IMPLEMENT-ATION
SÃO PAULO	*REORGANIZATION OF THE SCHOOL NETWORK*	*Reorganization of the school network,* including: separation of physical networks, a new pedagogical-administrative model with increased work load, and greater distribution of resources.	Since 1995
SERGIPE	*EDUCATIONAL REFORM*	*Educational reform,* including actions aimed at providing greater space for public schools, creation of pedagogical projects and democratization of schools.	Current
TOCANTINS	*ASSOCIATIONS AND AGREEMENTS*	Activities aimed at *improving both the quality of life and the quality of education,* using associations to create projects: cooperative schools, shared management, long-distance education (Fique Ligado).	Current

Source: CONSED (1996)

CHAPTER 11

BUILDING A STATE EVALUATION SYSTEM: THE EXPERIENCE OF THE STATE OF AGUASCALIENTES, MEXICO

Margarita Mª. Zorrilla Fierro[1]

Since 1992, following the implementation of the national decentralization process, the Mexican state of Aguascalientes has been developing a system for evaluating education. This article briefly describes its origins, evolution, current characteristics, and problems. It also analyzes the initial results obtained and the relationship established between the various state, national, and community organizations with regard to the evaluation of academic learning.

INTRODUCTION

The State of Aguascalientes has little more than five years of experience in educational administration, having previously been dependent on the Federal Government of Mexico. During this brief time, the state had to face the challenge of a shift to a more open and participative educational system that is accountable to the community, which is both its beneficiary and its principal resource. The state encountered cultural, governmental, and technical problems when it introduced an easy-to-understand educational administration system that took into account previous achievements and shortcomings. Previously, evaluation had been considered a form of political control: results of evaluations of the educational system's performance were kept confidential and were not published or used to understand or improve the system.

The experiences of building a state educational evaluation system in Aguascalientes suggest that results may be achieved in a relatively short period of time as long as certain institutional circumstances and leadership exist. The thoughts of participants in this effort suggest that:

- an evaluation policy may have unwanted effects that must be considered in advance;
- the evaluation actions require institutional responsibility;
- technical support from universities and/or research centers is essential;
- policies with respect to disseminating the information should be established at the outset;
- the teachers and, even more, the supervisors, need training;
- the participation of the users (teachers, students, and parents) in the process is a key element for the success of evaluation policies; and
- learning from the experiences of others prevents errors and helps to clarify alternatives.

THE STATE OF AGUASCALIENTES

Aguascalientes is a small state located in the center of the Republic of Mexico. Its population is less than one million. It has changed, over the past eighteen years, from a rural to an industrial society with an economy dominated by companies involved in metalworking, textiles, and, very recently, information sciences.

Its basic educational system (preschool, primary, and secondary) is responsible for approximately two hundred thirty thousand students, who constitute 1 percent of the National educational system.

When the National Agreement for the Modernization of Basic Education was signed by the Federal Government, the state governments, and the National Teachers Union (SNTE) on May 18, 1992, a new phase in the federalization process began. Educational services in Aguascalientes and the other states were no longer totally dependent on the Federal Government. The federative entities took charge.

This process, and an expansion of basic education, coincided, in Aguascalientes, with a new State Government effort to make education a priority. This new-found interest in education can be explained by the convergence of several factors:

- new political will on the part of the government;
- the presence of local expertise, specifically, education professionals and research and development experts, with links to both the national and international educational communities, who had been trained over the previous fifteen years at the Autonomous University of Aguascalientes or UAA;
- awareness of the reality of the educational system;
- the educational authority's capacity for coordination and negotiation;
- favorable geographic conditions with respect to size and physical access; and
- cultural characteristics favoring innovation.

In 1992, because of the need to fully understand the state of the basic education system the Federal Government was handing over, an initial evaluation of effectiveness and equity indicators was performed (IEA-CETE, 1993a). This led to three conclusions: the state ranked between ninth and fifteenth (out of a total of thirty-two) in relation to the

indicators assessed (see Table 1); the effectiveness of basic education in Aguascalientes had remained stable over the previous twenty years; and the condition of the educational system within the state was far from uniform, with significant differences being noted among educational levels, locations, municipalities, rural and urban zones, and schools with different types of financing (IEA-CETE, 1993b).

Table 1. Indicators of Effectiveness in Aguascalientes: Primary Education[2]

	School and Grade Enrollment Indicators	Definition	Applicable Educational Level(s)
1	Students in first grade of primary school WITHOUT preschool	Proportion of students enrolled in first grade of primary school who did not attend preschool	Primary
2	Grade Repetition	Proportion of students reregistered in a grade level	Primary and secondary
3	Overage Students	Proportion of students older than the established age for studying at a certain grade level	Primary and secondary
	Indicators of Attention to Students		
4	Students per group	Proportion of students per group or grade	Preschool, primary, and secondary
5	Students per teacher	Proportion of students per teacher	Preschool, primary, and secondary
	School and Grade Completion Indicators		
6	School Year Effectiveness	Proportion of students who complete and pass a grade level	Preschool, primary, and secondary
7	Effectiveness of teaching, reading, and writing	Proportion of students who complete the fourth grade of primary school with respect to those who started four years earlier	Primary
8	Level completion effectiveness	Proportion of students who finish an educational level with respect to those who entered that level the year before	Primary and secondary
9	Failure Rate	Proportion of students who do not pass a grade level	Primary and secondary
10	Coefficient of absorption into secondary school of those who have finished primary school	Proportion of students who finish primary school and enter secondary school	Primary-secondary link

Source: State of Aguascalientes. (1993). "Educación Básica. Diagnóstico de Indicadores de Eficiencia y Equidad."

Against this background, several challenges were identified:

- coordinate the three levels of basic education;
- provide a quality basic education to which all children would have access;
- ensure that students remain in school and complete each grade on time;
- make the teaching relevant to the students' lives;
- resolve problems by designing multidimensional strategies;
- identify the characteristics of each school in order to obtain quality results in each unique educational institution; and
- include the teachers, principals, and supervisors in developing solutions.

ORIGIN OF THE STATE EDUCATIONAL EVALUATION SYSTEM[3]

The development of the state evaluation system was influenced by two main premises: that there is an overriding need for the educational system to be responsible for the results it produces, and that information on the quality of learning is needed in order for this to occur. Both the measurement of various educational outcomes and the value judgments that are made on the basis of those measurements represent fundamental elements of micro and macro decision making policy leading to improving the quality of education.

For the first time in many years, an educational policy framework was established in Aguascalientes. This widely disseminated framework, known as the State Education Plan 1992-1998, asserted from the beginning the need for evaluation and its link to strategic educational planning. It should be noted that in Mexico—and Aguascalientes is no exception—evaluation is viewed as a way to persecute and settle scores. And especially in basic education, evaluation is seen as a way for those in authority to "show teachers up."

The groundwork for building a state system of educational evaluation was laid in 1983, when Don Mario Aguilera Dorantes[4], in agreement with the Independent University of Aguascalientes, performed a set of diagnostic studies of the educational system. The performance of these studies and the decision to disseminate their results—at least among school directors and supervisors—contributed to this distinguished teacher becoming responsible for basic education and teacher training in Aquascalientes. However, the political environment of those years was not at all favorable toward academic change. Section I of the SNTE was controlled by a group more interested in defending its sinecures than improving education. The material derived from these studies was preserved until, more than ten years later, the same teacher, Aguilera—then the president of the National Technical Council of Education—published a book containing the study results. However, the information never reached Aguascalientes.

The relationship between the UAA and the Aguascalientes Educational Institute (IEA)[5] goes back to 1978, when the University created two degrees in education, one directed toward educational research, and the other toward educational psychology consulting. From 1981 to 1983, a master's program in educational research was developed with

financial and technical support from the Delegation of the Department of Public Education in the state, the institution which preceded the IEA. Beginning in 1983, the relationship between the UAA and the state organization completely broke down, and it was not until 1992, when the new state government began the decentralization process, that academic specialists in education gathered to produce the state development plan. Both the Governor of the state and the Secretary of Education showed an interest in establishing links with the UAA from the beginning of their administration, as demonstrated by the fact that the educational system development work team was drawn mostly from academics, and by an agreement to replicate part of the 1983 studies evaluating learning. (This study was carried out in 1993; the results indicated that students had not improved in comparison to ten years earlier and, in some grades and subjects, had even gotten worse.) For his part, the current chief of the UAA has encouraged links between the University and the educational sector in various ways, such as through the performance of university studies of the quality and evaluation of education.

In 1994 and 1995, tests to evaluate academic learning were carried out in a sample of public and private primary schools. Publication of the results for the private schools caused discomfort, and even tension, between these schools and the IEA. At that time the educational authority decided not to publish the public school results, which were worse than the private school results, because of the potential negative impact on the public schools. Public schools teach more than 90 per cent of the students at the primary level and, furthermore, were declining because of system expansion policies and because of neglect of their development and quality.

Another immediate antecedent of the state evaluation system, although less well-known, is the evaluation of academic improvement that occurred within the framework of the Teaching Career Program.[6] This new system, which dates from 1993, uses a horizontal scale to measure teachers' professional performance. This system is significant because it is the result of a compromise between the Department of Public Education (SEP) and the SNTE that resulted from negotiations established by the aforementioned National Agreement for the Modernization of Basic Education (1992).

The evaluation of educational outcomes, especially from school learning, was added to the national educational system in the context of the Compensatory Programs being implemented since 1991 in several entitites in Mexico, and which are currently in effect in twenty-four of the thirty-two states. Aguascalientes does not have this type of program because it deems that its grade repetition rate is statistically insignificant. Despite this, state evaluation policies have been enhanced by the national experience with the compensatory programs.

With this as background, the state educational authority made a qualitative leap by implementing, in the 1995-1996 school year, state examinations for all students completing the sixth grade of primary school and the third grade of secondary school. Furthermore, by sampling, they measured the results of learning in the third, fourth, and fifth grades of primary school and the first and second grades of secondary school.

The intent was to design and operate a system to evaluate the quality of education through various educational outcome indicators. This would permit the state, municipality, education zone, and school level authorities to identify support needs and design intervention strategies to improve the quality of education.

The concept of quality of education includes assuring all children and youths the provision of educational services, entry, and continued enrollment in school, including timely completion of each educational level, and the knowledge acquisition that is relevant to the students' present and future lives.

The measurement of learning in school is perhaps the most important component, but not the only one. An evaluation system encompasses other types of quantitative and qualitative information from various indicators. When these are considered together, the improvement in the quality of education can be assessed.

In March 1997, the Education Law of the State of Aguascalientes, which established the authority and responsibility of the Aguascalientes Educational Institute to evaluate the educational system, was enacted. It states:

> **Article 18.** The Aguascalientes Educational Institute shall design and operate a State Educational Plan that shall take into account all educational types, levels, and modalities. It shall consider both the quantitative and qualitative aspects of present and future educational activities. The State Educational Plan shall be supported by the diagnostics and evaluations of the State Educational System. This shall have a system of educational information that shall contain updated information on students, teachers, and schools, as well as potential demand based on demographic and economic dynamics.
>
> **Article 19.** With the same purpose and in harmony with the actions and powers of the Department of Public Education established in the General Education Law, the Aguascalientes Educational Institute shall develop and coordinate a State Evaluation System for education that shall take into consideration all the components of the State Educational System and the quality dimensions and indicators established by this Law. To ensure that the evaluation carried out shall be rigorous in content and methodology, the opinion[s] and support of teachers and education professionals shall be obtained.
>
> The individual evaluation and assessment of the students, for purposes of accreditation and certification of studies, shall be accomplished separately in accordance with applicable federal and state regulations.

With this regulation, the evaluation of education is permanently established in law and beyond the reach of governmental administrations.

EVALUATION OF LEARNING IN 1996

The evaluation experience of 1996 stands out because of its magnitude. State examinations were administered to all students in the final grades of primary and secondary school (sixth and third, respectively), students in the third, fourth, and fifth grades of primary school, and the first and second of secondary school in a sampling of schools.

Examinations were administered in mathematics and Spanish, except in the sixth grade, where the natural sciences and history were also tested.

Coverage was follows: 21,400 students in the sixth grade of primary school; more than 21,000 students in the third, fourth, and fifth grades of primary school; 13,800 students in the third grade of secondary school; more than 8,000 students in the first and second years of secondary school from a sample of 25 per cent of the schools. Finally, the students who entered the first year of secondary school in 1996–1997 also received a test of academic skills known as the Academic Diagnostic Instrument for Students Entering Secondary School (IDANIS).

This statewide evaluation was carried out in four stages: test design, test administration, statistical analysis of the results, and their communication and dissemination.

The test instruments were designed by basic education teachers who are specialists in the subjects covered and educational researchers from the Autonomous University of Aguascalientes.[7] Their development was based on the curricular contents of the 1993 Primary Education Plan and Curricula. Although the Spanish and mathematics tests for both primary and secondary school were inspired by Spanish instruments, they were based strictly on the Mexican plan and curricula for each grade. These tests consisted of fifty items each.[8] In addition, the Spanish test included a question requiring the students to develop a written text that could be used for subsequent research on writing [skills]. The tests had different items for each grade in accordance with the respective curriculum, although they were identically structured in terms of the central thematic ideas. The natural science and history tests for the sixth grade had fifteen items each.

The tests were administered during the second and third weeks of June. Teachers in each Basic Education Zone (ZEB)[9] were rotated so that no teacher would administer the examination to his own group or in his own school. In addition, a team of individuals not involved in the process observed the testing.

The results were analyzed, and the total score for each test in each grade was broken down at different levels and with different variables. Overall analyses were performed using averages, dispersions, and comparisons by subject, as well as the student's sex, the school's environment (rural/urban), type of school (public/private), session (morning/evening), municipality, and ZEB.

The feedback process was initiated by the team from the IEA's evaluation area before data complication and analysis were completed. Strategies for disseminating and

communicating the results included giving each child in the sixth grade of primary school a certificate with his average test score; developing a document with the averages by school and subject for internal circulation; circulating press releases by the educational authority; and presenting the results at meetings of principals and directors and at local and national fora on educational research and evaluation.

SOME RESULTS OF THE MEASUREMENT OF THE INDICATOR OF SCHOLASTIC PROGRESS IN PRIMARY SCHOOL

The results suggested that most students do not attain the expected level of academic performance, and that performance tends to be better in Spanish than in the sciences and mathematics. Deterioration in the average results is also noted as the students move from one grade to another, particularly in mathematics (Graph 1). Significant differences were noted between the results in rural and urban environments, with urban environments' results being better; and between public and private schools, with private schools having the advantage.

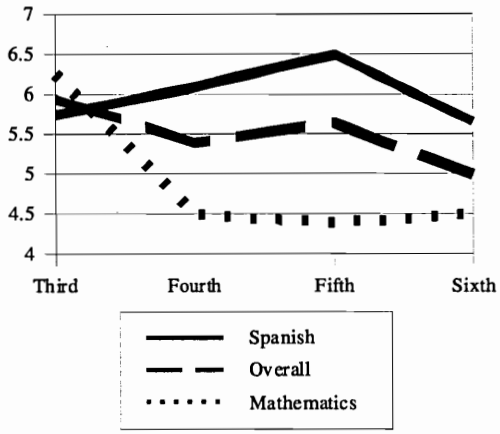
The differences between boys and girls were not significant; nor were the differences between public schools operating in morning and evening sessions, or between schools in the rural part of the state capital's municipality and those in the rest of the municipality.

The same trends were noted in the sixth grade. Because of our experiential knowledge about the operation of the educational sector, particularly the ZEBs, and about the schools, we know that the aforementioned differences are largely due to the type of attention the students, teachers, parents, and schools receive. For example, the results for the capital's municipality, the most urban region in the state, where more than half the population is concentrated, are relatively better than in the rest of the municipalities. However, the municipality of Cosío (ZEB "Q"), located in one of the poorest areas in the state, achieves results just below those of the capital. Their results are even better than those of the Rincón de Romos (ZEB "V") municipality, a prosperous community with a long teaching tradition. This situation is due, in large part, to the fact that, in Cosío, the supervisory team is accomplishing an important task, while Rincón de Romos has not been able to rely on solid academic leadership. Similarly, the results for the municipality of Calvillo (ZEB "O-P"), which reflected the worst indicators of educational effectiveness in 1991, stand out. This improvement in its learning outcomes is also due to the supervisory team's work.

These results emphasize the need to strengthen pedagogical management. The school staffs' and areas of supervision's identities become more meaningful in the context of the new decentralized administration.

However, the historical entrenchment and political configuration of school supervision and direction raise serious questions and challenges in terms of reorienting these managerial and supervisory functions toward administration governed by professional criteria.

Graph 1. Comparison of overall averages and course averages by grade levels[10]



Source: State of Aguascalientes. (1996). "State System for Evaluation of Education. Report of Results."

The results for primary and secondary education are identically distributed. The averages reflect the same trends of deterioration. What other hypotheses can be proposed to explain these results? We would like to point out some cultural arguments. On Teachers Day[11], the Secretary of Education of Aguascalientes spoke to the teachers, saying:

> Why do some states in this country have better educational results than others? Could it be, perhaps, that in this cultural mosaic that is our country, certain communities have trusted in their education while others have not? It seems that there is something in this. Aguascalientes, despite its favorable conditions of size and communication, reached the nineties with educational performance indicators around the national average, even lower. I venture to say that the community of Aguascalientes, in the recent past, attributed little importance to its education; it did not pay attention to it. Our industrialization is very recent and, therefore, the demands it brings are also very recent. When our life was culturally rural, we did not need academic education, the demands were fewer, life was simple. Now, the scenarios have changed radically and in a very short time, thus the growing demand that children and youths receive a quality education, that they really learn about matters relevant to their present and future lives, as people and as citizens.
>
> I believe we all agree that academic education is the fundamental agent of culture and, if culture is the factor that makes the difference between one community and another, the conclusion seems obvious. (Álvarez, 1997)

The cultural hypothesis goes further, transcending and permeating all the machinery of the educational sector: the initial training of teachers, the constant refresher training of

active teachers, school conditions, the school's relationships with the community and parents, the pedagogical methods used by the teachers, the value of knowledge, the relationships among the various educational agents and actors[12], the cultural traditions of the various regions of the state and country, the mechanisms for access and continuity in school supervision and direction, the complicated union situation, and others. These thoughts represent the starting point for a more in-depth investigation of the *modus operandi* of each of these dimensions of ordinary school operations. This is the question: what are the characteristics of the educational supply in a certain place and how do they interact with the characteristics of demand and context to produce certain educational results?

EFFECTS OF POLICIES ON EVALUATION

The policies implemented in Aguascalientes have been relatively successful in that they have not been impugned either by the community or by the teachers union, as has traditionally occurred. One reason has been the formation of a technical team of educational research professionals, but a large part of this initial success was determined by the will of the educational authority, the Secretary of Education. The same could be said of other states. When there is genuine political will to give attention to evaluation, one can shape a supply of information and judgment about education despite enormous difficulties.

But work on the demand side is also fundamental. In Aguascalientes, disseminating the results for children in the sixth grade of primary school was also a factor. The state educational authority had the courage to circulate the evaluation results. As previously mentioned, each student received his results in a personalized way, with a certificate. In the press, the results were disseminated through bulletins from the Secretary of Education. Union leaders spoke in favor of evaluation and the need to improve the quality of learning. The Governor of the state gave prizes to the one hundred top students in the state, and ten of these were sent to meet the President of the Republic.

The tests results were presented in greater detail in the summer of 1996, at the meeting of state authorities for basic education,[13] and the various comparisons were highlighted. The initial reaction was denial, or at least minimization, of what the results demonstrated. Real concern was generated by the finding that the existence of a number of schools and teachers adequate to serve the population is necessary but not sufficient. To comply with the constitutional precept of guaranteeing education to all individuals is, above all, to guarantee the right to learn, and to have opportunities for development. Those in the educational system who are in positions to direct and make decisions strongly resist the evaluative process.

Moreover, administering examinations to evaluate students' academic progress is an activity external to the school but not the educational system. The agency responsible for performing the evaluation is the Headquarters for Educational Development of the selfsame Aguascalientes Educational Institute. However, in 1983, the evaluation was external to the schools and the educational system itself. It was the universities that

performed it. Nothing is worse for teachers than feeling that they are being shown up by others whom they see as totally foreign to the educational system's problems.

Despite the lack of specific studies on the impact of these evaluative activities, we can say that the community has lost its fear of reporting on the state of children's and youths' learning. Parents, one way or another, know their children's level of learning, or at least sense it. Disseminating the results corroborates their suspicions about the deterioration of the quality of education provided in the schools.

To place the students' learning at the center of the debate, of the analysis, and of the work itself is a Copernican type of revolution because it now permits the teachers, parents, and the students themselves to pursue another way to learn. The centrality of the student in the educational process is slowly becoming reality. One element which, in our judgment, mitigated the impact of the poor results on the teachers was the initiation, in January of 1996, of an extensive refresher training program for all active basic education teachers. For the first time in more than thirty years, the teachers were benefitting from innovative course and workshops strategies. From the start, the political impact on the teachers was favorable; in their own words, they feel cared for, "taken into account."

The effect of evaluation from the teachers' perspective is variable. When it is a question of broadened attention and continuing training for the teaching profession, the evaluation results are discussed; however, critical self-analysis still has no place in the teachers' discussions. They continue to say that the results are due more to cultural or socioeconomic factors affecting the students than to the schools' operating mechanisms, of which each teacher is, in the final analysis, a part.

The results of the evaluation at the end of the 1996–1997 year may contribute information on progress or regression. It is important to mention that those in the central administration of the educational system are uncertain about improvements in students' learning. It seems that the inclusion of innovations like the refresher training program and the new curriculum are demanding new knowledge and pedagogical skills from the teacher. In fact, these require time—we do not know how much—to be assimilated and incorporated into daily teaching practices. In other words, intellectual discourse and conviction need to be processed and applied before they can generate new pedagogical practices in the school and classroom.

THE RISKS AND ALTERNATIVES

In our experience, in the Mexican educational system, innovations tend to become perverted quite rapidly. It is important to point out the possible undesirable effects of an educational evaluation system. These include: that the schools teach only for purposes of passing the state examinations; that evaluation is a function that the teachers repudiate as attributable only to the system's central authorities; and that the results are used to stigmatize the students, to pressure them irrationally, or even to expel them from the schools.

With respect to the teachers, the same thing can happen: the teacher is stigmatized by control without support; savage competition among the schools and the teachers; and peoples' balanced and total development is hurt more than helped. For the moment, we still do not know if the curriculum is appropriate, and yet evaluation of the students' learning is already linked to salary increases in the horizontal scale framework promoted by the Teaching Career Program.

To reverse this type of unwanted effect, we must be able to work in several directions, namely:

- Continue measuring the scholastic progress indicator in the sixth grade of primary school and the third grade of secondary school, and add the third grade of intermediate education.
- Decide on the concept and meaning of a state educational evaluation system and its place in a new organizational structure which is moving from isolated compartments to a matrical distribution of tasks, which are resolved by consensus of all the responsible authorities. In this sense it is necessary to put the problems in the hands of those who are interested in thinking about and implementing the solutions.
- Promote educational research through agreements with the higher education institutions which perform it. Give attention to matters such as designing and testing instruments and methodologies to evaluate scholastic centers, training, and strengthening work groups of educational administration experts, and developing evaluative studies that frame the results in the context of educational supply and demand.
- Establish institutional policies for circulation and dissemination of results, and orient and diversify development programs such as the one for continuing refresher training for the teaching profession.
- Give priority to training supervisors, directors, and teachers in educational evaluation, identifying the responsibilities of each in his field of competence. The evaluation must become a means of support, a mechanism for accounting for actions, and a path to growth and learning for children, youths, and adult educators.
- Understand and learn from other experiences based on the same evaluative research and other countries' systems. This is obligatory if we are to avoid repeating the same errors.
- Comprehend the political component of each evaluative activity. It is not just a question of the will of whoever directs the educational system. Rather, evaluation must be at the center of consensus among educational agents and actors. As indicated in the last Informational Letter from the IIPE-UNESCO (Ross, 1997), research on educational results and policy about where and how much to invest in education is a *complex mixture*. The most important lesson is that a dose of prudence is needed when drawing policy conclusions from the reports of educational results, if these have not been preceded by detailed studies of the context and conditions in which they are produced.

LESSONS FROM THE FIRST EXPERIENCE

This briefly described experience teaches that it is possible to build state evaluation systems. In particular, we should point out the factors that favor it and that we consider fundamental.

The first factor we consider fundamental is the decision and political support of the State Education Secretary, which translated into the conviction that evaluation of learning is a necessity; the decision that evaluation must be external to the school but not to the educational system; the allocation of budget resources; the supervision and follow-up of the process; and the backing of the responsible team.

Second was a trained technical team of professionals with experience in educational research, in performing evaluative studies, teaching basic education, and in computer science. The state team is composed of ten people, of whom three direct the processes of developing and validating tests, designing the data gathering operation and training, and developing different types of reports of results for various users. This group has developed the ability to administer the evaluation system by incorporating the national and state[14] approaches to educational policy into its design and operation.

Third was the state educational authority's courage to publish and disseminate the results to the community in general and to various users in the state educational system in particular.

Another factor which, in our judgment, has been very important is the refresher training for primary and secondary teachers. As has been mentioned, all the teachers are being served by an innovative training plan for active teachers. This occurs through workshop courses—which take place on business days and during working hours, with teachers and materials that travel from one location to another—for groups of teachers who belong to a basic education zone. This program was initiated six months before the first administration of the state examinations.

The fifth factor we identified is the confidence and respect that the team from the Headquarters for Educational Development and from the evaluation area has been able to garner during the last four years through the tasks they perform in refresher training, research, innovation, and the evaluation itself.

This element came to light this year. On the one hand, tests were being administered to students (from the third grade of primary school to third grade of secondary school) of the teachers who are enrolled in the Teaching Careers Program. This evaluation of progress is included in a national system centrally controlled by the SEP itself. Unfortunately, this evaluation has been discredited, and has generated undesired effects that have perverted this mechanism. On the other hand, the majority of teachers perceive what we call state examinations as something original with and belonging to Aguascalientes, administered more to know where the school system is rather than to rate the teachers or show them up.

Until last year (1995–1996), the administration of the examinations for the Teaching Career indicator of scholastic progress was carried out by another agency of the Aguascalientes Educational Institute. This year—in which these examinations are being administered for the fourth time—is the first time the Evaluation Area is responsible for doing this, making it possible to spot problems, deficiencies, deviations, and fakery.

Finally, in the aforementioned policy address by the State Secretary of Education, he insisted that evaluation is key to improving the quality of education and asserted that the poor results—attributable to both the structure and operation of the state educational system—constitute the starting point. Although the students' scholastic progress is a reflection of their teachers' work, it is clear that it is also the reflection of the performance of the various levels of authority: of the school principal, the school supervisor, and the authorities and mid-level managers of the Aguascalientes Educational Institute.

NOTES

[1] Director of the Headquarters for Educational Development of the Aguascalientes Educational Institute and teacher-researcher for the Department of Education of the Independent University of Aguascalientes. Like all intellectual works, this cannot be conceived as the exclusive product of the signer. What is presented here is the integration of various texts that document the policies with respect to evaluation of education in the State of Aguascalientes. It emphasizes the work on analysis of the results of various measurements of academic learning performed by Professor Daniel Eudave and his collaborators. The comments, always analytical and constructive, of Pilar González, currently Director of Planning of the Aguascalientes Educational Institute, have enhanced this modest document. Responsibility for what is stated herein is solely the author's.

[2] As can be seen, these indicators of effectiveness do not include measurement of learning outcomes.

[3] Zorrilla, F. M. Ed. (1997) *Descentralización e Innovación Educativa. Una Mirada Desde el Proceso de Gestión de la Innovación. El Caso de Aguascalientes.* This work was financed by the Ford Foundation and coordinated by the College of Mexico.

[4] At that time he was the Delegate of the Department of Public Education in the entity. Professor Aguilera is prominent in the national educational system. He entered as a rural teacher in 1924 and was twice Chief Clerk of the same Department.

[5] In the state, the management of the education system is the responsibility of a decentralized public organization, the Aguascalientes Educational Institute. The highest authority is the General Manager, whose functions are similar to those of a secretary of education.

[6] [The] Teaching Career [Program] is a teacher performance evaluation system on a horizontal scale on which the students' academic progress, among other factors, is evaluated. This evaluation affects teachers' salaries.

[7] The Department of Education of the UAA has developed, among other lines of educational research, an evaluation of scholastic progress in the areas of knowledge as well as skills and has three researchers who are experienced in the field. With respect to the evaluation of learning that occurred in 1996, the UAA participated with personnel who worked on the development of the Spanish and mathematics tests for the third through sixth grades of primary school. The natural sciences and history tests for the sixth grade of primary school were developed by primary school teachers with the help of UAA researchers. The Spanish and mathematics tests for the three secondary school grades were developed by teachers of said level. The reproduction and distribution of the tests were the responsibility of the IEA.

[8] Those for Spanish and mathematics were valid in the sense that the majority of the items had an acceptable capacity to discriminate and level of difficulty.

[9] The entity's elementary education services are distributed among geographic units known as Basic Education Zones, or ZEB. These zones comprised preschool, primary, and secondary schools, physical education services, and special education. They are served by a team of supervisors directed by a coordinator. In total there are twenty-two zones, twelve in the capital's municipality and one in each municipality for a total of ten; they are of different sizes with regard to the number of schools, teachers, and students attending. For more information, see Márquez, M. and Zorrilla, M. (1997) *Redefinir la Supervisión para Atender la Escuela Singular.*

[10] This document presents the results obtained by students in the third through sixth grades of primary school in the State of Aguascalientes on the state examinations administered in June 1996. This time of year was chosen, because by that time most of the curricular content in the evaluated subjects has been covered. A stratified sample was taken for the selection of schools to evaluate in the third through fifth grades. The state was divided into strata according to the following criteria: urban and rural environment; socioeconomic condition and social well-being: favorable, unfavorable, and marginal. For the sixth grade of primary school, the tests were universally administered.

[11] In Mexico, on May 15, when Teachers Day is celebrated, schools and cities are full of festive activity in homage to teachers.

[12] A distinction between agent and actor in terms of education that can be comprehensive and fruitful is the following: the actor receives a script and acts it out, while the agent acts, decides, and interprets.

[13] The State Secretary of Education, all the mid-level managers of the IEA, the coordinators of the basic education zones, and some of those responsible for projects met.

[14] The General Education Law, the Federal Government's Programa de Desarrollo Educativo 1995–2000 (Educational Development Plan 1995–2000), the Education Law of the State of Aguascalientes, and the State Education Plan 1992–1998.

REFERENCES

Álvarez G., Jesús (1997). *Mensaje del Día del Maestro. Mayo de 1997.* Mecanograma.

IEA-CETE (1993a). *La Educación Básica en Aguascalientes, Diagnóstico y Propuesta.* UID: Aguascalientes.

IEA-CETE (1993b). *Programa de Modernización de la Función Supervisora. Documento Rector.* Aguascalientes, 87 pp.

Ross, Kenneth N. (1997). "Investigación y Política: Una Mixtura Compleja." *Carta Informativa del IIPE,* XV, 1, 1-4.

CHALLENGES AND POLICY OPTIONS FOR EDUCATIONAL EVALUATION

Benjamín Álvarez H. and Ray Chesterfield

Educational evaluation is a subject of strategic importance for nations. Both national social policies and international agreements reflect the interest shared by all countries in achieving the highest quality of learning for their children and youth. The agreement signed at the World Conference on Education for All held in 1990 in Jomtien, Thailand, illustrates the growing demand for improved educational systems throughout the world. Although there is ample consensus on such need and numerous educational reform efforts are being made, we have little information to help us judge their success or adapt educational systems to the needs of various types of students.

This book has described the evolution of and recent trends in educational evaluation and discussed innovations introduced as result of educational reforms that have taken place during the current decade in Latin America. These trends suggest new dilemmas and policy options that countries, states, and families will find themselves obliged to address. This chapter highlights certain of these tendencies, summarizes the status of current practice, and seeks to contribute to the policy debate.

PRINCIPAL TENDENCIES

Measurement of academic achievement

Following international trends, efforts to determine the performance level of students in basic skills such as language and mathematics are increasing dramatically in Latin America. The most widely used instrument for measuring such performance has been scholastic achievement tests. Until recently, emphasis was on tests based on statistical norms of the kind that are exceedingly common in the United States. This is partly because many countries have not yet established criteria for the development of basic skills in the grades or courses making up the system. In some countries, the evaluation of learning in order to increase the quality of education is predicated on the prior experience of testing systems designed to screen candidates for admission to higher levels of schooling.

All of the theoretical and practical works presented recognize the importance of using additional instruments to assess the results of academic learning, and of measuring other

important variables related to academic achievement. Some even describe alternative approaches and instruments designed to relate student performance to curriculum development or environmental factors. In other words, there is a trend that seeks not only to identify the results of the learning achieved but also to explain those results in order to implement strategic interventions. Nevertheless, broader aspects of student behavior (e.g., the formation of attitudes and values that constitute the objectives of many educational reform programs) are generally not included in evaluation systems.

Added to the interest in measuring academic performance is a move to institute national standards for academic performance. In addition, there is a desire to establish international agreements and comparisons aimed at improving the academic performance of students in each country. These agreements and comparisons will be based on the learning obtained through systematic and international interaction. Nevertheless, countries require a sound national information and research infrastructure in order to fully participate in such activities.

Definition of educational quality

Although the preceding chapters refer to various concepts of educational quality, most of the systems described appear to define it in terms of the internal deficiency of the system or from the standpoint of the productive function of education, by comparing it to productive enterprises.[1] The assumption is that the scholastic resources and processes established in a given school or system determine the degree of learning to be achieved. In other words, the proper combination of inputs may reduce waste in the system and increase system efficiency.

One related definition of educational quality that generally appears in discussions concerning the rationale for evaluation is based on the concept of external efficiency, i.e., the capacity of the products of an educational system to function and compete in society and to increase the ability of a nation to participate in the global economy (Heyneman, 1997). However, in selecting an input-process-outcome scheme to orient actions aimed at improving the quality of education, it is wise to bear in mind that this approach always involves only partial representations of processes upon which numerous factors exercise influence.

The evaluation systems discussed devote less attention to educational quality as the creation of an adequate environment for cognitive development and skills acquisition. This concept is based on recent work suggesting that learning takes place in an environment of collaboration that is based on familiar concepts and supported by a network of social interactions. Such collaboration allows students to interpret rather than accumulate information (Gardner, 1991 and Levinger, 1996). Given the emphasis of many reform efforts on constructivist approaches to learning, this perspective for viewing quality deserves consideration.

Rather than producing an abstract and finished concept of quality, this trend toward identification of educational settings that are more stimulating and that facilitate a

higher level of learning achievement—which is a subset of education quality—is producing a change process based on the relationships and actions of a community of educators, parents, politicians, administrators, and individuals affected by education (Wenger, 1996). The development of processes of participation, organization of consortia, teamwork, and ongoing monitoring of academic performance constitutes the basis for creating a national community of learning related to education.

Value added

What does the school or educational program add to student learning? This question attempts to understand evaluation as a function of the concept of "value added." The established goals and the learning outcomes of a particular educational system or school are affected by a number of factors, one of which is the academic performance of students prior to their enrollment in the respective program or school. For example, judging the work of a school based exclusively on the scores obtained by its students on final tests, with no consideration of their context or the initial performance of the student population, could lead to errors in interpreting the work of teachers and create misconceptions about the relative strengths or weaknesses of various school systems.

Teacher performance

Teachers are considered to be central players in the educational process. It is assumed that once they achieve an appropriate level of education or training they will be able to carry out their functions satisfactorily. The low results of established indicators of educational efficiency in many countries of the region call this assumption into question.

Several countries have observed that it is essential to have available support mechanisms for the professional development of teachers, such as opportunities for training and systems for evaluating their professional performance. The studies and cases presented in this book suggest close ties between the performance of teachers and that of their students. Yet the evaluation of the professional performance of the teacher based on the measurement of the academic performance of his or her students is not in itself sufficient to improve the system as a whole. Several countries of the region feel that it is necessary as well to establish criteria and forms of evaluating the practice of teaching, despite the fact that past efforts have met with little success.

There is, in addition, a need to ensure active teacher participation in educational reform efforts and to establish professional standards. As pointed out in the discussion on Colombia's New School program and Chile's 900 Schools program, making teachers participants in the design of reform activities can be successful in terms of improving the quality of education. These cases also suggest that such participation by teachers leads to reflection on their work and to a sort of self-evaluation. These experiences and the application of models such as those presented in the "Practice Series" in the chapter on teacher evaluation suggest the potential for overcoming the traditional resistance of teachers unions to evaluation efforts.

Emphasis on consortia

Considerations involving the established technical capacity of countries, the costs involved in providing educational services, and the changing role played by the states have increased the number of consortia and partnerships both inside and outside the educational system. These partnerships, which have been set up for purposes of conducting educational evaluation, comprise a wide variety of players. In several countries, for example, such consortia involve collaboration among ministries of education, universities, and private research centers. In such cases, some components of the evaluation process are entrusted to organizations within the educational community but external to the ministry of education because they may perform certain tasks with greater efficiency.

A second type of consortia involves the participation of a country in international evaluation programs. As suggested in the chapter on curriculum evaluation that includes the description of the TIMSS project, comparisons of academic performance among countries may help identify differences in the objectives, processes, and content of national programs and contribute to the development of reform policies. Participation in such programs provides experience in new evaluation methodologies that may be applied in specific countries. However, countries with fewer resources should carefully consider the advantages to be obtained from their participation, the extent to which their systems are comparable to those of other countries, and the way in which the resulting information would be used.

Regional efforts to develop common standards represent the third type of consortia described in this work. Such consortia help countries learn from the experience of others and develop a consensus about elements critical to any process of educational reform, such as success indicators or evaluation mechanisms. They also enable countries to acquire knowledge of the "best practices" that may inspire innovations in other contexts, reveal the types of student work considered to be of high quality, or shed light on the way in which information problems are solved. But such consortia require long-term commitments and policy support at both the national and international levels.

Lastly, the studies presented stress the consortium between government—through ministries of education—and civil society. This occurs when the debate over educational reform is opened to the individuals involved at all levels and in all areas of the educational system. Although these debates have focused on the efficiency of the system rather than on its results in terms of learning, in most countries the problem of standards to which education must respond begins to take on increasing importance in the arena of public opinion.

Local capacity

The strengthening and use of local capacity to perform evaluation tasks are closely related to the development of institutional exchanges. This capacity is made up of three fundamental elements: a critical mass of researchers and analysts, efficient institutions,

and information networks. Available data indicate that in several countries of the region there exist both a human and institutional capacity and a remarkable experience in developing and implementing broad-based systems for evaluating academic achievement. The problem lies in strengthening that capacity, expanding the range of activities in order to fuel a process of social learning aimed at improving education, and linking the critical mass of evaluators and researchers through a series of national and international information networks. For those countries having less experience, efforts are made to create local organizations and learn from the experience of others.

Increase in the use of information

All of the works presented reflect a concern for the use of the information produced by evaluation systems. Although the potential users of such information form a broad spectrum of groups, including parents, students, teachers, administrators, politicians, and researchers, there is but limited evidence of the way in which these individuals assimilate and use the results produced by evaluation systems. To a degree, the situation presented here is similar to that involving the use of social research in general, despite the fact that the results of evaluation tend to have a more direct impact on interested parties.

Although the cases analyzed present and discuss several occasions when the results of the evaluations or studies were used in an almost linear and direct fashion in decision making, the process by which knowledge permeates society and is subsequently understood and adopted tends to be broader and less predictable. For instance, it appears to be an incremental and accumulative process that makes use of multiple channels. As interest in education increases and civil society participates more directly in decisions involving education, the evaluation of education results requires increased dissemination and the use of several different information channels. The studies presented here point out that the information produced by evaluation systems is beginning to be distributed through the media and other strategies as well as through reports.

IMPLEMENTATION OF EVALUATION SYSTEMS

Comparison of the case studies in the preceding chapters provides information useful for identifying the conditions and processes needed to organize an educational evaluation system. These experiences may illuminate the policy decision of countries ready to consider the development of broad-based programs for monitoring student academic performance.

In countries with a more recognized tradition of educational research and a more cohesive institutional infrastructure, national evaluation systems tend to be more stable, serve a broader clientele, and offer greater potential for using the information produced than in countries with a more recent history in this area. The creation of the technical experience, institutional capacity, and networks of contacts is a lengthy process, but of considerable import for the development and use of systems for monitoring education and its results. The existence of research centers and university-based post-graduate programs facilitates this process.

All of the countries of Latin America and the Caribbean possess a basic—albeit limited—system for compiling information on general indicators of educational development. Due to the emphasis placed on facilitating access to education for greater numbers of children, available indicators refer to rates of enrollment, grade repetition, and school dropout, and include some information on physical infrastructure. Few national education systems are able to measure academic performance beyond the evaluations conducted by teachers in the classroom.

The experience of the few countries that have developed national evaluation systems—initially for screening purposes and subsequently to improve the quality of education—suggests that there are certain basic requirements for the initial operation of such systems. The principal requirements are technical capacity and political consensus. The stability and continuity of such systems depends on a series of factors that include, in addition to technical resources, management capability, the presence of institutions whose mission is to conduct evaluation and research, and the development of information and communication systems.

The region's limited experience with programs of international evaluation and monitoring indicates that such programs provide a decisive impetus to local programs. Exposure to existing programs helps countries learn about the operation and effectiveness of various alternatives. From this standpoint, the role of international agencies may be critical, as such organizations can link different players and provide access to unlimited resources for local or national activities. However, such programs require a solid base in the corresponding countries, with clearly established institutional responsibilities, technical capacity, and political backing.

POLICY OPTIONS

The expectations created with regard to the effects of educational reform tend to be quite high. It is expected that the reforms will contribute to solving long-standing problems, such as poverty and inequality, and assist in increasing the competitive capacity of nations. In most countries, however, the reforms are being implemented amidst economic adjustments, shortages of resources, and downsized government structures. This context poses constant dilemmas for social policy and decisions related to establishing priorities for public expenditures. Although the creation and strengthening of the capacity for critical judgment and evaluation in the system as a whole emerge as logical consequences of reform programs, it is essential to clarify what their contribution is and how their significance, use, and technical quality will be ensured. The most urgent policy topics related to the above-described trends are described below.

Should investments be made in educational evaluation?

Most countries of the region lack the experience and infrastructure required to create broad-based evaluation systems. Instituting such systems involves high costs and their results are not immediately visible. One initial policy consideration is whether the investment in the organization of educational evaluation systems is justifiable or

whether the already established custom of exams administered by teachers in their respective courses will suffice to determine the outcomes of education and identify those responsible for its successes or failures. In this case, the question regarding the achievement of "national standards" could not be answered properly, nor would it be a significant issue. The assumption would be that an investment in other areas more directly related to teaching would improve education more. The argument to support this position would be that the existence of evaluation systems does not necessarily guarantee improved education in a given setting.

But the option of investing in the creation and adaptation of mechanisms designed to generate information about the operation and results of educational systems can be rationalized. Such rationales include the need to link society as a whole to school management, the ability to collect useful information that could permanently inform the decisions that both government and citizens need to make, and the urgency to define our own social utopias and the ultimate purpose of national education systems. Moreover, the social dynamic created by evaluation constitutes one of the principal engines of the reform process, since critical policy issues are debated, and new directions outlined, around the information they produce and the issues they stir up.

How should quality in education be measured?

As previously mentioned, the type of student performance measured depends to a large extent on the implicit or explicit definition that a country may have of the quality of education. If interest lies in broad measurements of system operation, indicators on schooling will be appropriate, though not sufficient, to determine achievement. If quality is conceived in terms of the results of learning, a testing system will be appropriate. Standardized or norm-based tests constitute one option that has been used in a number of countries of the region. Case studies show, however, that there is a growing interest in ensuring that all students attain a minimum level of competence in language and mathematics in primary school, thus suggesting the development of performance standards and the use of criterion-referenced tests.

As mentioned in several chapters, none of these ways to address the issue of quality in education responds to the concept of quality as an environment for promoting critical thinking. A common criticism of standardized tests is that they penalize imagination and critical thinking, as they reward those students who quickly select the answer considered "best" by the test designers. Complementary measurements, such as those known in the literature under the name of "authentic evaluation" that include student portfolios or work files, exhibitions, observations, and other manifestations of educational achievement, are quite appropriate. (This technique may be part of teacher training programs that address evaluation of school learning.)

"Quality of education" is a heuristic concept that will likely evolve as our knowledge of learning and available technology develops and evolves. Evaluation systems facilitate such conceptual developments and experimentation with alternatives. In addition, they allow partial achievements to be recorded in the ongoing search for improved educa-

tional opportunities that can be put into practice through an iterative process between knowledge and action.

What other aspects of evaluation should education policies take into account?

The evaluation systems in place in most countries focus on measuring academic achievement. But, as suggested in several of the preceding chapters, there are other components of the educational setting that need to be considered in policies that govern system monitoring, such as the processes that supposedly lead to outcomes (what happens in the classroom, what is taught); the performance of the professionals responsible (teachers and directors); and school organization at both the central level (education ministries and secretariats) as well as at the periphery (the schools).

The teacher is a central player in the education process. Evaluation of his or her professional performance and contribution to student learning has been a controversial issue. The obstacles to developing procedures for evaluating teachers identified to date include several of a theoretical nature: what exactly constitutes good teaching, for example, is not always easy to explain. Other concerns are employment or labor-related, such as concerns over job stability or the loss of labor union power. Still others involve problems concerning professional ethics and justice. Investing in the professionalization of teachers or their evaluation or in monitoring their task constitutes a policy dilemma.

Within the context of a shortage of resources, an additional dilemma surfaces: whether to evaluate students or teachers. As the task of teaching becomes professionalized and schools acquire increased autonomy, teachers themselves will be more involved in the task of evaluating their own efforts. New instruments based on sound ethical, professional, and research-based criteria are opening new horizons for resolving the dilemmas created by existing policies.

The studies cited in several chapters of this book suggest the importance of using indicators of the classroom educational process to complement the evaluation of education's effects. Such information contributes greatly to the public debate, the development of curriculum policies, and the proper orientation of the textbook and educational materials industry.

Investment in consortia

Consortia have considerable potential as a way to support evaluation systems. In the national context, consortia between the ministry of education and other educational support organizations can aid in developing and implementing educational evaluation more efficiently, by making the best possible use of the advantages offered by each. International consortia provide an opportunity to learn about new technologies and build common operational frameworks among nations. Such investments can be extremely positive from the cost-benefit standpoint. The consortium between the state and civil society, when based on a broad consensus as to the need for monitoring and

evaluation, can aid in overcoming the traditional obstacles by which the issue has been affected, such as contoversy over the evaluation of the teaching profession.

Consortia and partnerships are not, however, without cost, as they involve not only the cost of participation but also the costs incurred in managing the relationships among members, selecting participating organizations and individuals, and complying with the obligations deriving from the commitments made. International programs that compare academic achievement risk stressing competition, disregarding analysis, and diminishing national support for education vis-à-vis the results obtained, when those results do not meet expectations. These and other issues related to benefits possible from participating in this type of consortia should be given consideration in each case.

Investment to promote the use of the information from the evaluation

It is important to know what evaluation systems can and cannot do. They cannot, by themselves, improve education, nor do they have the ability to reveal all of its results. Likewise, they cannot directly resolve the problems affecting the system. They can, however, provide significant information over time that will make it possible to compare systems and results, detect problems, and provide data and analysis for policy decisions. They represent an essential instrument for implementing appropriate change, developing new alternatives, and involving the whole of society in the debate on education.

International experience in the use of the information for social change in general, and for changes in education in particular, has shown that neither information nor knowledge, in and of themselves, lead to social change. Policies and decisions in the field of education are not guided primarily by either research or evaluation. Policy formulation is an ongoing process that includes institutional and interpersonal relationships, multiple information sources, and negotiation processes. The influence of information on policy decisions depends on numerous factors, such as the context in which the information is presented, who presents it, when it is presented, and the importance and quality of the presentation.

There is growing recognition that, if significant changes are to be made, information must be distributed, debated, and analyzed in such a way as to obtain a shared meaning among the professionals involved on the various levels of the educational system. Such a process of developing knowledge alters concepts and may lead to behavioral changes inspired by those concepts. This suggests that teachers, parents, administrators, and decision makers in the educational system must be involved in the development, analysis, interpretation, and use of knowledge related to teaching and learning.

One policy alternative is to invest in information dissemination. Such dissemination should not be aimed exclusively at communicating to the actors decisions that have already been made. Instead, the information disseminated should enable them to participate in the creation of shared meanings, new concepts, interpretations, and information. This can lead to a system that includes strategic plans for involving

stakeholders in the use of the knowledge generated and in the production of new knowledge about how to expand learning opportunities in society. The ability of information technology to monitor the education of each individual, the performance of systems, and the activities of various organizations and actors in the education arena are unparalleled in history. The challenge for evaluation and research is to create conditions that foster the ongoing learning of nations.

CONCLUSIONS

Educational reform faces two fundamental problems. The first is philosophical in nature: what are the purpose and desirable characteristics of education in each society? What are the social utopias that justify education for all, public schools, support for private education, and a social organization that facilitates an ongoing process of learning? The other problem is technological: how to plan and implement the necessary changes? The mechanisms of learning and renovation provided by evaluation respond to this dual concern.

Much of the technology necessary for the systematic evaluation of students, teachers, and schools is available in the countries of Latin America and the Caribbean. In addition, most countries have available a critical mass of specialists with the ability to successfully implement programs of educational research and evaluation; the countries also possess a limited institutional capability. The existence of such resources, however, does not ensure that evaluations will be conducted or that their results or the results of educational research will be used in policy decisions.

Once a decision has been made to invest in educational evaluation, countries can take a number of steps. These steps appear to contribute to ensuring the usefulness of evalutaion results in terms of improving the quality and effectiveness of education. They include, as a minimum, the following:

- The development of a consensus with respect to the evaluation and research agenda that meets the needs and concerns of a wide array of actors on the educational stage;
- The use of existing knowledge through a review of the international experience and the commitment of local organizations working in the areas of evaluation and research;
- The creation of consortia among the ministry of education and other organizations experienced in the fields of research and evaluation, with a view toward conducting joint endeavors;
- The development and use of a series of analytical instruments for understanding educational phenomena in order to enhance learning in children; and
- The use of evaluation findings to provide feedback that will lead to new conceptions and effective actions so that ongoing improvement and learning will become a substantial element of the culture of educational systems.

NOTES

[1] For a summary of studies of the productive function in education, see, for example, A. Rubin "Assessing Designs for School Effectiveness Research and School Improvement in Developing Countries," *Comparative Education Review,* vol. 41:2, 178-204, May, 1997.

REFERENCES

Gardner, H. (1991). *The Unschooled Mind. How Children Think and How Schools Should Teach.* New York: Harper Collins, Basic Books.

Heyneman, S. (1997). "Economic Growth and the International Trade in Education Reform." Paper presented at the USAID Conference on Human Capacity Development for the 21st Century, Washington, DC, July 14-18, 1997.

Levinger, B. (1996). *Critical Transitions: Human Capacity Development Across the Life Span.* Newton, MA: Education Development Center.

Wenger, E. (1996). *Communities of Practice.* London: Cambridge University Press.

APPENDIX

INTERNATIONAL PERSPECTIVES ON STANDARDS AND ASSESSMENT: A SELECTED BIBLIOGRAPHY

Teresa Kavanaugh

Airasian, Peter. 1993. "Policy-Driven Assessment or Assessment Driven-Policy?" *Measurement and Evaluation in Counseling and Development,* vol. 26.

The general purpose of policy-driven assessments is to produce educational change and to improve the performance of pupils, teachers, and the educational system in general. More specifically, two aims guide policy-driven assessment. First, assessments are used to improve teacher and pupil performance by heightening standards and motivating teachers and pupils to work harder. Second, assessments are used to gain control over the objectives of education and in some cases, over the process of teaching, in order to focus, clarify, and influence what is taught in classrooms and schools. This paper states that policy-driven assessments have an important symbolic effect: tests and assessments symbolize order, control, and desirable school outcomes. They are powerful moral symbols of a traditional set of social and educational values like hard work and reward for effort. A more accurate name for what has gone on in the name of educational reform in the prior decade, suggests the author, is assessment-driven policy, not policy-driven assessment. Airasian concludes that there is one discernible trend and one likely conflict that are related to future assessment policy. The trend involves efforts to broaden assessment from primary reliance on multiple-choice-item formats to increased reliance on authentic assessment. The conflict involves the level—local, state, or national—at which the most influential assessment will be carried out in the next decade.

Capper, Joanne. 1996. *Testing to Learn/Learning to Test.* Washington, DC: Academy for Educational Development.

Too often, educational tests and national assessments measure superficial learning (i.e., memorization of facts) rather than understanding and command of the concepts. This book is designed to be a comprehensive guide to improve educational testing in developing countries. It addresses the relationship between examinations and assessments, and teaching and learning, in these countries. While chapters two through seven focus

on the more technical aspects of test design and implementation, chapters one and eight provide valuable insight into issues related to testing and assessment at the national level.

Greany, Vincent, and Kellaghan, Thomas. 1995. "Equity Issues in Public Examinations in Developing Countries." World Bank Technical Paper No. 272, *Asia Technical Series.* Washington, DC: World Bank.

This study presents an analysis of inequities associated with public examinations in developing countries. Research from close to thirty countries, mainly in Africa and Asia, is reviewed. Because of the high stakes attached to examination performance, teachers teach to the examination and, as a result, opportunities for students who leave school at an early age are inadequate. Exam-related practices that may create inequities for students are presented, including scoring procedures, the use of culturally inappropriate questions, the requirement that candidates pay fees, private tutoring, examination in a language with which some candidates are not familiar, and various malpractices. The report notes that using quota systems to deal with differences in performance associated with location, ethnicity, or language group membership also creates inequities for some students. The authors conclude that the limited available evidence does not indicate that examinations create inequities between genders. Further, they state that ranking schools on the basis of students' examination performance may not provide a fair assessment of the work of schools.

Greany, Vincent, and Kellaghan, Thomas. 1996. *Monitoring the Learning Outcomes of Education Systems.* Washington, DC: The World Bank.

For many years, governments have collected and published statistics on how their education systems are working and developing, particularly focusing on indicators such as school numbers and facilities, student enrollments, and efficiency indices such as student-teacher ratios and rates of repetition, dropout, and cohort completion. However, few countries have systematically collected and made available information on the outcomes of education. This book focuses on monitoring the learning outcomes of educational systems as a means of improving the quality of student learning. It provides an overview of the nature and choice of outcome indicators; a review of national and international assessments; information about national assessments and public examinations; suggested components for a successful outcome-based national assessment; and a fictitious case study of a national assessment, with the objective of reviewing the guidelines previously outlined in the book. The book distinguishes between the use of outcome indicator types, such as those that measure students' cognitive and affective development, and reviews the differences, advantages, and disadvantages between national and international assessments. Contrasting national assessments and public examinations, the book offers guidelines for stakeholders critical to ensuring a successful national assessment, asserting that not only should stakeholders be in agreement, but that the most significant and influential stakeholder is the ministry of education.

Greany, Vincent, and Kellaghan, Thomas. (Undated). "The Integrity of Public Examinations in Developing Countries." Mimeo.

The focus of this paper is to examine the procedures under which tests are prepared, administered, and scored, and how they are observed or violated. The report stresses that only if such procedures are successfully implemented will the integrity, wholeness, and soundness of examinations be maintained. The proper implementation of procedures ensures that no candidate is placed at an advantage or disadvantage relative to other candidates because of unfair practice; it verifies that the marks or grades awarded candidates are directly related to the ability that is being measured rather than to irrelevant factors or uncontrolled conditions. The study reviews the sources of evidence available on the topic of malpractice and the forms of malpractice that have been identified in examinations. It considers reasons for malpractice and outlines some procedures that have been developed to detect it. Finally, the report explores ways to control or prevent its occurrence.

Horn, Robin; Wolff, Laurence; and Velez, Eduardo. 1991. "Establecimiento de Sistemas de Medicion del Rendimiento Academico en America Latina: Un Analisis de los Problemas y la Experiencia Mas Reciente." Washington, DC: World Bank. Latin America and the Caribbean Technical Department, World Bank Paper No. 9.

In this paper the authors outline how to design and administer tests and review important factors that should be taken into consideration in this process. The experiences of Chile, Costa Rica, Mexico, and Columbia are cited as successful experiences in national assessment. However, although these countries have had successful national assessments, none of them have incorporated components dedicated to the research and development of the tests themselves. The report emphasizes that to guarantee that tests actually lead to better quality in education, countries should consider the importance of defining long-term objectives for assessment goals and creating a plan for disseminating test results.

Kellaghan, Thomas, and Greany, Vincent. 1996. "Using Examinations to Improve Education: A Study in Fourteen African Countries." World Bank Technical Paper No. 165. *Africa Technical Department Series.* Washington, DC: World Bank.

This synthesis report describes the types, functions, performance levels, governance, administration, and funding of public examinations based on a set of studies undertaken on primary and secondary examinations in fourteen African countries. Procedures for funding examinations; constructing, administering, and scoring papers; and reporting results in each country are outlined. While the primary function of public examinations in these countries is to raise academic standards and select students for the next level of the educational system, findings of this report reveal that while these examinations may help raise academic standards, they may also help give rise to problems in the education system.

Madaus, George, and Kellaghan, Thomas. 1991. "Student Examination Systems in the European Community: Lessons for the United States." Contract Report submitted to the Office of Technology Assessment. Washington, DC: United States Congress.

In the first part of this report, the evolution of testing policy is reviewed, and six proposals to establish national exams in the United States are considered. In the second part, the origins and development of public examinations in Europe are analyzed. The authors describe the complexity, operation, and contexts of the examination systems of European countries, focusing particularly on France, Germany, and the UK, and outlining the three major features of each country's public examination systems. Looking at their selective function, the major part played by universities, and the role of examinations in defining student learning—what it is students learn and how they learn—the paper considers the implications of these examinations for American schooling.

Mislevy, Robert. 1994. "What Can We Learn from International Assessments?" Paper presented for the Conference on the Use of International Education Data, Washington, DC.

International assessments have been thought of as yielding information that allows comparisons of relative achievement by county and subject, or that allows the improvement in one country from the determinants of achievement in another, or finally as a way to provide information to policy makers on the status of achievement and practices in their own countries. In this paper, the kinds of inferences that can be drawn from international educational assessment are explored, considering the evidence that can be obtained and how it can be interpreted. It is argued that indices of educational achievement that are to varying degrees comparable across nations can be useful, but that ascertaining the relative standing of nations will tell very little about how to create educational policy or to improve instructional practice.

Murphy, Paud; Greany, Vincent, and Lockheed, Marlaine. 1996. *National Assessment: Testing the System.* Washington, DC: World Bank (Economic Development Institute).

This book is designed to provide policy makers and educational practitioners working in developing countries with information about different aspects of national assessment systems and how they work in practice. It outlines a number of different features of national assessments for individuals interested in studying such systems. Designed to encourage a discussion of the quality of education provision in countries, how it should be measured, and how to use the information gathered to improve quality, this book is written primarily for policy makers and practitioners in ministries of education. However, it is also intended to be of value to those seeking to gain information about various technical aspects of national assessments and how these systems function in developing countries.

O'Neil, John. 1993. "Can National Standards Make a Difference?" *Educational Leadership,* v. 50, no. 5.

This article reviews and examines issues and progress-to-date in the national standards movement in the United States and concludes by outlining some concerns raised by those who criticize national standards. It cites the work of the National Council on Education Standards and Testing as key to clarifying goals and definitions of national standards. The Council called for a standards-based education system monitored by a national system of student examinations. Specifically, the Council recommended that national standards be developed, and that they include content standards (what students should know and be able to do); student performance standards (the level[s] of student competence in the content); and system performance standards (to assess the success of schools, districts, states, and the nation as a whole in helping all students attain high performance standards). In addition, the Council said that states should develop school delivery standards to judge whether schools are providing students with the opportunity to attain high standards. The council also recommended that a national assessment system linked to these standards be developed.

Operations Evaluation Department, World Bank. 1994. "Building Evaluation Capacity," *Lessons and Practice.* Http://www.worldbank.org/html/oed/lp004.htm#evaluation. Washington, DC: World Bank

This paper looks at the steps needed to build and to benefit from evaluation capacity in the public sector in developing countries. It outlines ways in which evaluation can play a critical role in four areas of a nation's public sector management: 1) influencing policy analysis and formulation; 2) improving resource allocation and the budgetary process; 3) improving investment programs and projects; 4) examining fundamental missions of institutions or the government itself. Outlining current problems with evaluation in developing countries, the paper notes that many developing countries still lack the essential requirements for effective evaluation: the quality of information and access to it is often poor, mechanisms for feedback into the decision making process are weak, and a culture of accountability is not firmly in place. The report provides lessons for planners, emphasizing that evaluation can be best developed if it is seen by all concerned—within both the country and the development community—as a way to learn and to improve the performance of the public sector. It reviews how to develop a country-specific strategy and provides observations on developing evaluation capacity.

Ravitch, Diane. 1997. *National Standards in American Education: A Citizen's Guide.* Washington, DC: Brookings Press.

Advocates of national standards claim that clear and consistent standards would improve academic achievements and prepare the nation's students to deal with the challenges of the 21st century. Opponents feel that standards would be too homogenous and unresponsive to the culturally and economically diverse population and that they would give too much power to the federal government to mandate what their children should be taught. This book explores both the promise of a nationwide system of

standards and the problems surrounding their implementation. By highlighting the U.S. experience and drawing comparisons to other countries, the author summarizes the case for and against a national system of standards and assessments, as voiced in the context of the U.S. debate. She then describes the debate in terms of why the consensus for reform materialized and how the momentum for change grew.

Resnick, Lauren, and Nolan, Kate. 1995. "Where in the World are World-Class Standards?" *Educational Leadership*, vol. 52, no. 6.

Countries known for their outstanding students have several practices in common; clear, consistent, demanding public education standards head the list. In an effort to come up with a definition of world-class standards, the New Standards Project at the University of Pittsburgh began its international benchmarking efforts in 1993, with the hope of collecting and analyzing the standards documents of other countries. Trying to get a clear picture of world-class standard performance, the authors examine tracking, curriculum, and exams in the United States, Netherlands, France, Sweden, and Germany. They learned that the shared practices and common threads among different approaches to education in these countries teach important lessons: setting clear, consistent, demanding, public standards helps students perform well; tracking and grouping practices must make sense in the culture of the school and in view of both the student's and community's future goals; exams should test what students have been asked to learn, preferably in the same ways they must perform in class; exams that call for complex, demanding tasks can be given to a wide range of students, perhaps to all students. As front-line professionals in the education process, teachers should have much to say about what goes into exams and how they are graded.

Sanders, William, and Horn, Sandra. 1995. "Educational Assessment Reassessed: The Usefulness of Standardized and Alternative Measures of Student Achievement as Indicators for the Assessment of Educational Outcomes." *Educational Policy Analysis Archives*, Vol. 3, No. 6.

Methods of assessment based on the use of standardized tests have come under intense fire in recent years with some critics going so far as to call for their complete elimination. This paper presents the debate about whether standardized or alternative assessment is a better model for evaluating educational outcomes. Central to the debate is the determination of which educational indicators are best suited to assess whether students have achieved the goals set out for them. The paper states that standardized tests render viable, inexpensive, reliable, and valid indicators of student learning that are particularly useful in the assessment of educational entities and student achievement. The authors stress the importance of alternative forms of assessment as viable tools for measuring student progress and achievement as long as special attention is given to ensure their validity and reliability. The paper concludes by advocating the use of multiple indicators of student learning, including those provided by standardized tests.

Schiefelbein, Ernesto. 1993. "The Use of National Assessments to Improve Primary Education in Chile." In *From Data to Action: Information Systems in Educational Planning,* edited by David W. Chapman and Lars O. Mahlick. Paris: UNESCO.

Although three national assessments of student achievement have been undertaken in Chile in the last 20 years as a basis for improving educational quality, these efforts have neither generated the expected increments in students' achievement nor increased equity among students from different socioeconomic groups. This paper describes the aims, design, operation, and impact of the three programs. A major finding was the remarkable stability in fourth graders' achievement scores over time, interpreted by some as evidence of stagnation of the education system. This case study explores possible causes of such stability, suggests improvements in the use of test results, and examines the appropriateness of the analyses used to makecomparisons between groups of schools and across years.

Snyder, Conrad. 1997. "Exam Fervor or Fever: Case Studies of the Influence of Primary Leaving Examinations on Uganda Classrooms, Teachers, and Pupils." Washington, DC: Academy for Educational Development.

In many education systems, examinations define success and failure for individuals, act as gatekeepers to future opportunities, and attach credibility to the systems that engage them. Uganda's Primary Leaving Examinations (PLE) exemplify these features. This report addresses several key questions that explore the influence of the PLE on Uganda classrooms, teachers, and parents. Given the general influence of examinations, how do they affect the children and adults who are involved with them? Can the specific nature of the examinations influence the way education is delivered? Are there simple strategies that can manipulate the impact of examinations on the classroom? The report concludes that: 1) examinations in Uganda have substantial effects on the individuals involved; 2) examinations do influence the way education is delivered; 3) the impact of examinations is not simple. The report concludes by stating that national examinations are complicated components of a complex system. Although national assessments provide a lever for central policy control over instruction in the classroom, their use in reform is not as simple as it might first appear.

The World Bank. 1995. *Priorities and Strategies for Education: A World Bank Review.* Washington, DC: World Bank.

Chapter six of *Priorities and Strategies for Education: A World Bank Review* highlights the importance of educational objectives in setting priorities and strategies for education. Arguing that insufficient attention has been placed on educational outcomes thus far, whether determined by the labor market or in learning terms, the chapter discusses how outcomes can be used to set and monitor public priorities, focusing particularly on rate of return analysis. Issues of setting standards and monitoring performance in education, issues which come into play when the public sector in a given country has made decisions concerning the allocation of public resources, are also addressed. The book states that once standards for performance have been set, performance needs to be

observed and evaluated. The most common way to do so, say the authors, is by both tests and examinations. The chapter concludes by outlining the policy and pedagogical applications of public examinations and testing.

CONTRIBUTORS

Benjamín Álvarez H., Ph.D., a graduate of the University of New Mexico and member of the International Academy of Education and the American Evaluation Association, is a Senior Policy Analyst with the Academy for Educational Development in Washington, D.C.

Thomas Kellaghan is the Director of the Educational Research Center at Saint Patrick's College in Dublin, Ireland. A graduate of Queen's University in Belfast, he has authored and coauthored twenty books and monographs and one hundred twenty specialized articles. He is a member of the International Academy of Education and the European Academy.

Gilbert Valverde is Assistant Professor of Measurement and Quantitative Methods at Michigan State University and the Associate Director of the United States National Research Center in charge of overseeing the country's participation in the Third International Mathematics and Sciences Study (TIMSS). He holds a doctorate from the University of Chicago.

María Inés Cuadros Ferré is a psychologist on the faculty of Saint Xavier's University in Colombia. She coordinated the efforts of the American Secretariat ProTempore for the Monitoring of Agreements on Childhood from 1994 to 1996 from her posts in the Office of the Colombian President and the National Planning Department. She is presently working for Save the Children Fund, U.K.'s South American Program.

Erika Himmel, a specialist in educational evaluation and measurement, is a professor on the faculty of the Catholic University of Chile. She headed up the team that developed the screening system for students applying to Chilean universities and helped set up Chile's Educational Quality Assessment System, known as the SIMCE.

Gabriel Restrepo is a professor of sociology on the faculty of the National University of Colombia. He has served as President of the Colombian Sociology Association and as head of the Social Development Unit attached to the National Planning Department.

Carol Anne Dwyer is the Executive Director of the Division of Program and Educational Policy Research at Educational Testing Service (ETS) in the United States. She is the principal designer of the Praxis Series: Professional Assessments for Beginning Teachers, a national series of teacher examinations offered by ETS.

Francisco Álvarez is a researcher with the CIDE, the Center for Educational Research and Development, in Chile. He holds a degree in education and is an expert in teacher training. Paula Vergara and María José Álvarez are also on the CIDE research team.

William Webster is the Director of the Research, Evaluation and Information Systems Division for the Dallas public school system. He holds a doctorate from Michigan State University. He has published over fifty specialized articles on evaluation and applied statistics.

Robert L. Mendro, a Ph.D. from the University of Colorado, heads the Research Unit for the Dallas public schools.

María Teresa de la Fuente is a consultant on the project for the strengthening of educational quality in the State of Paraná, in Brazil.

Heloisa Luck, who holds a doctorate in education from the University of Columbia, is a consultant for the Paraná State Education Department.

Corina Lucía Ramos holds a master's degree in education from the Federal University of Paraná. She is presently working as a consultant in the area of educational management planning and auditing.

Margarita Zorrilla is the Director of Educational Development for the Aguascalientes Education Institute, Aguascalientes, Mexico. She is also a professor on the faculty of the Autonomous University of Aguascalientes.

Ray Chesterfield is Vice President of Juárez and Associates. He holds a Ph.D. in international education with a specialization in anthropology. His work has focused on the creation and evaluation of programs that promote the formation of human resources in marginalized communities.

TITLES IN THE TECHNICAL PAPER SERIES

1. *Paths of Change: Education Reforms Under Way in Latin America and the Caribbean. / Senderos de Cambio: Génesis y Ejecución de las Reformas Educativa en América Latina y el Caribe.* Benjamín Álvarez H. and Mónica Ruiz-Casares, Editors/Editores. 1997.

2. *Partnership for Change: Using Computers to Improve Instruction in Jamaica's Schools.* Errol Miller. 1996.

3. *Evaluation and Educational Reform: Policy Options. / Evaluación y Reforma Educativa: Opciones de Política.* Benjamín Álvarez H. and Mónica Ruiz-Casares, Editors/Editores. 1998/1997.

OTHER TITLES ON POLICY REORM FROM THE ABEL 2 PROJECT

1. *Education Reform Support.* Luis Crouch, Joseph DeStefano, and F. Henry Healey. 1997.
   Volume 1: Overview and Bibliography
   Volume 2: Foundations of the Approach
   Volume 3: A Framework for Making it Happen
   Volume 4: Tools and Techniques
   Volume 5: Strategy Development and Project Design
   Volume 6: Evaluating Education Reform Support

2. *An International Curriculal Perspective on Decentralization: An Introduction to its Problems, Prospects, and Evaluation.* Richard Kraft. 1995.

3. *Improving Capacity for Policy Analysis and Planning in Cambodia's Ministry of Education, Youth, and Sports.* Christopher Wheeler, Kay Calavan, and Melinda Taylor. 1997.

4. *Primary Education for All: Learning from the BRAC Experience.* Colette Chabot, Manzoor Ahmed, Rohini Pande, and Arun Joshi. 1993.

5. *BRAC's Non-Formal Primary Education Program: A Customer-Focused Evaluation of the World's Largest NGO.* Anne T. Sweetser. Forthcoming.

6. *Framing Questions, Constructing Answers: Linking Research with Education Policy for Developing Countries.* Noel F. McGinn and Allison Borden. 1995.

7. *Teacher Development: Making an Impact.* Helen Craig, Richard J. Kraft, and Joy du Plessis. Forthcoming.

To request copies of publications, a complete catalog, or further information, contact:

ABEL 2 Project
Academy for Educational Development
1875 Connecticut Avenue, NW
Washington, DC 20009-1202

Telephone: 202-884-8000
Fax: 202-884-8408
E-mail: abel@aed.org

For more information or additional copies of this publication, contact:
ABEL Clearinghouse for Basic Education
Academy for Educational Development
1875 Connecticut Avenue, NW
Washington, DC 20009-1202

Telephone 202-884-8288
Facsimile 202-884-8408
E mail abel@aed.org
Internet www.aed.org/intl/basiced.html