Developing Educational Assessment Systems
in Latin America

A Review of Issues and Recent Experience

by

Robin Horn, Laurence Wolff and Eduardo Velez

Human Resources Division

September 1991
DEVELOPING EDUCATIONAL ASSESSMENT SYSTEMS IN LATIN AMERICA

A Review of Issues and Recent Experience

By

Robin Horn, Laurence Wolff
and Eduardo Velez'
There are two ways educational systems can be organized. One is to keep teaching children without feedback on what students learn. The other way is to institute a national assessment system to monitor student learning for feedback and improvement.

Although most educational systems in the world today follow the first approach, the situation is changing. Many countries, including several in Latin America, have realized the potential of measuring student achievement for the purpose of improving the quality of education. Chile, Costa Rica, and Mexico are countries where such monitoring takes place with potential tangible results towards a better educational system. Education programs in Belize, Brazil (Northeast and Sao Paulo), Colombia, Dominican Republic, Ecuador, and Jamaica include proposals to establish educational assessments.

Implementing educational assessment is not an easy task. If this is to be done properly, it requires technical expertise, financial resources and institutional commitment. This report distills experience with educational assessments in LAC and other countries for the purpose of arriving at suggestions and recommendations on what should be done and what should be avoided when designing a new system.

It is hoped that this report will prove useful to the increasing number of countries in Latin America which are now preparing or considering the institution of an educational assessment system.

George Psacharopoulos
Senior Human Resources Adviser
Technical Department
Latin America and the Caribbean
ABSTRACT

The objective of national education assessment systems is to evaluate the progress of school systems in achieving curricular or other goals of the education system and to identify problem areas in curriculum and instruction. Key decision points in the process of implementing and using educational assessments include: defining objectives and costs, establishing an assessment management system, selecting the appropriate type of tests to be given, designing and applying effective instruments, disseminating test results and providing feedback, supporting complementary research, and overall ensuring that assessment results feed into the decision making process at both local and national levels. Many countries in Latin America have begun to implement educational assessments. The experience in Chile, Costa Rica, Mexico and Colombia is of particular interest. To ensure that assessments in fact do influence policy makers as well as individual teachers to improve the quality of education, Latin American countries should especially consider defining, from the start, long term objectives and costs of the assessment system; designing, from the start, and then implementing dissemination and feedback systems; contracting much of the work to agencies outside government; using sample surveys, where appropriate, rather than censuses; and supporting research on policy issues raised by assessments.
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPEIP</td>
<td>Centro de Perfeccionamiento, Experimentación e Investigaciones Pedagógicas (CHI)</td>
</tr>
<tr>
<td>DiSCuS</td>
<td>Diversified Secondary Curriculum Study</td>
</tr>
<tr>
<td>DGEIR</td>
<td>Dirección General de Evaluación y de Incorporación y Revalidación (MEX)</td>
</tr>
<tr>
<td>EAEPPS</td>
<td>Evaluación del Aprendizaje en Educación Prescolar, Primaria y Secundaria (MEX)</td>
</tr>
<tr>
<td>ECIEL</td>
<td>Estudios Conjuntos sobre Integración Económica Latinoamericana</td>
</tr>
<tr>
<td>ETS</td>
<td>Educational Testing Service (USA)</td>
</tr>
<tr>
<td>IEA</td>
<td>International Association for the Evaluation of Educational Achievement</td>
</tr>
<tr>
<td>IIMEC</td>
<td>Instituto de Investigaciones para el Mejoramiento de la Educación en Costa Rica</td>
</tr>
<tr>
<td>IRT</td>
<td>Item Response Theory</td>
</tr>
<tr>
<td>MOE</td>
<td>Ministry of Education</td>
</tr>
<tr>
<td>PER</td>
<td>Programa de Evaluación del Rendimiento Escolar (CHI)</td>
</tr>
<tr>
<td>SECE</td>
<td>Sistema de Evaluación de la Calidad de la Educación (CHI)</td>
</tr>
<tr>
<td>SEP</td>
<td>Secretaría de Educación Pública (MEX)</td>
</tr>
<tr>
<td>SIMCE</td>
<td>Sistema de Información sobre la Calidad de la Educación (CHI)</td>
</tr>
<tr>
<td>SNEE</td>
<td>Sistema Nacional de Evaluación Educacional (MEX)</td>
</tr>
<tr>
<td>SNP</td>
<td>Servicio Nacional de Pruebas (COL)</td>
</tr>
</tbody>
</table>
TABLE OF CONTENTS

SUMMARY AND RECOMMENDATIONS ........................................... i - vi

I. INTRODUCTION .............................................................. 1
   Issues in Primary Education in Latin America ....................... 1
   The Purpose of this Report ........................................... 2

II. THE OBJECTIVES OF ASSESSING STUDENT LEARNING ACHIEVEMENT ... 3
   Definitions ......................................................... 3
   How Educational Assessments Can Be Used .......................... 4
   Using National Assessments in Educational Research and Development ...... 7

III. TECHNICAL AND MANAGERIAL ISSUES IN ASSESSMENT SYSTEMS ...... 9
   What and When to Measure .......................................... 9
   Designing an Effective Instrument ................................ 12
   Administrating the National Assessment ............................ 17
   Disseminating Test Results and Providing Feedback ................. 20
   Managing the Assessment .......................................... 23
   Expanding the Scope of a National Assessment System ............... 25
   Cost .................................................................... 27

IV. ASSESSMENTS SYSTEMS IN LAC: STATE OF THE ART .................. 29
   Chile--History and Background ....................................... 29
   Chile--Technical Aspects of the Assessment System ................ 30
   Costa Rica--History and Background ................................ 32
   Costa Rica--Technical Aspects of the Assessment System .......... 33
   Mexico--History and Background ..................................... 34
   Mexico--Technical Aspects of the Assessment System ............... 35
   Colombia--History and Background ................................... 36
   Colombia--Technical Aspects of the Assessment System ............. 37
   The LAC Experience: Summary and Recommendations ................ 38

REFERENCES ................................................................... 41

ANNEXES

1. Sample Feedback from Kenya ........................................... 43
2. Sample Research Results from Colombia, Peru and Brazil .......... 44
3. Sample Background Questions for ECIEL Study .................... 47
4. Sample Questionnaires from Colombia’s Entrance Exams .......... 49
5. Sample Feedback from Chile and Costa Rica ....................... 51
6. Sample Colombia Guidelines on the Methodology Used in the Secondary Entrance Exams (Information for Teachers) ................ 54
SUMMARY AND RECOMMENDATIONS

The purpose of this review is to alert education decision-makers in Latin America to the range and complexity of the issues involved in developing and implementing national educational assessment systems. The report summarizes best practice in implementing and utilizing educational assessments, reviews the experience of several Latin American countries in implementing national assessment systems, and provides some suggestions to help ensure that educational assessment systems in Latin America can influence policymakers and teachers to improve the quality of education. This report is timely because many countries in Latin America and the Caribbean are currently embarking on programs to assess educational achievement in primary education. Given their economic problems and constraints on public finances, these countries need to direct their education sector resources towards the most cost-effective investments, programs, and interventions.

The availability of even the highest quality data from educational assessment systems cannot guarantee improvements in student learning in the region. This requires the commitment of national and local leaders, as well as teachers, school principals, and the general public, to learn, to understand and to be candid about educational assessment results, and to adjust schooling practices and education sector policies on the basis of information and analysis gleaned from national education assessments. This review provides suggestions to help ensure not only the effective implementation of educational assessments, but also their effective utilization to improve student learning achievement.

Objectives and Uses of National Educational Assessments

This report focuses on the design and use of national educational assessments. The objective of national assessments of student learning is to measure the educational performance of a nation’s or region’s students, to evaluate the progress of the schools, school districts, municipalities, or states in achieving curricular or other goals of the education system, and to identify problem areas in the nation’s curriculum, classroom instruction, and/or student behaviors. Educational assessments can be a cost-effective means of helping to improve learning outcomes. By itself, of course, measuring student learning will not yield increased student achievement any more than weighing grain will yield increased agricultural output. It is, however, a necessary condition to establishing quantitative targets, assessing the tradeoffs of alternative allocation strategies, input combinations, and instructional technologies, and allocating resources and effort to achieve established targets.

To improve educational quality by any means it is important to establish a base-line measurement of learning. National assessment systems can help educators identify effective inputs and processes, and improve these inputs and process to achieve measurable gains in student learning. Assessments can affect student achievement by mobilizing public attention and support for education, by providing information to educators who in turn revise instructional designs and teacher training, through rewarding, either directly or indirectly, good performance of teachers and schools, through contributions to the analysis of characteristic problems in students’ understanding and application of knowledge, and through associated research which identifies more cost-effective allocations of instructional inputs, more effective instructional processes and school environments.

To achieve the goals of improving learning through educational assessments requires a strong design and implementation effort. While it is obviously important to have adequate technical, financial, and institutional support, an often overlooked area is that of dissemination of the results of assessments. From the start, an educational assessment should include a detailed plan as well as adequate funds and staffing for dissemination. Another often ‘overlooked but key element needed to help educators and policy
makers pursue the goal of improved learning is the linkage of the assessment with a research program that measures the impact of various educational inputs on learning. Linking assessment information with subsequent labor market performance or other social outcome measures would also help national policymakers by gauging the economic rates of return associated with different educational objectives and outcomes. These types of research programs require collecting and processing data that can be used by the research community to conduct analyses, as well as building in adequate funding and establishing professional linkages with research institutions.

Technical and Management Issues in Educational Assessments

The following are the main issues and decisions which will need to be taken in the course of implementing educational assessments:

- **Defining Objectives and Costs.** The objectives of educational assessments determine what should be measured and how it should be measured. These objectives will therefore need to be clearly articulated from the start. Assessment objectives could include: (a) monitoring and reporting on the nation's progress towards established performance targets; (b) providing quantitative ammunition to encourage or to direct districts or schools to improve performance; (c) furnishing educators with data and research for the purpose of diagnosing and treating learning problems or of changing instructional design and teacher training; and (d) developing materials that encourage teachers to improve their instructional content and pedagogical practices. A decision will be required on what levels and subjects would be tested. A realistic five-year implementation and cost estimate will need to be prepared on the basis of the objectives and modalities selected.

- **Achievement vs. aptitude tests.** Achievement tests are examinations keyed to measuring the extent to which children learn the intended curriculum. Aptitude tests measure students' "innate" abilities. Achievement tests therefore would normally be used to serve the objectives of an educational assessment. Achievement tests have greater predictive power for future educational levels than aptitude tests.

- **"Objective" tests vs. performance tests.** "Objective" tests, such as multiple-choice or short-answer test formats, are more cost-effective, reliable, amenable to statistical analysis, and easier and quicker to score than performance tests, such as long essay examinations or skill demonstrations. However, if it is financially and technically feasible, then performance-type items could also be included in assessments to ensure the teaching and learning of writing, original thinking, analysis, and synthesis.

- **Criterion-referenced vs. norm-referenced tests.** Criterion referenced tests measure whether particular or prescribed standards are met. Norm referenced tests compare and rank student learning with average (mean) achievement levels. Criterion referenced tests are normally the preferred choice for assessments since they are designed to examine students’ mastery of the system's educational objectives. The performance standards of criterion referenced examinations will need to be made explicit and widely reported and understood.

- **Preparing an examination syllabus.** The preparation of a detailed examination syllabus for each subject area and grade-level to be assessed is the first step in instrument development. Objectives listed in the examination syllabus are normally distributed across all of the content, process, and context areas judged to be of fundamental importance. How much emphasis to place on recall of factual information and terminology, and comprehension of basic concepts and ideas, versus higher-order thinking skills, problem analysis, and the application of knowledge and skills will need to be decided.
Generating test and background items. Test items will need to be generated by trained item writers to correspond to the content and process areas identified in the examination syllabus. Assessment instruments will also normally measure a variety of student, classroom, school, and community input and cost items determined on the basis of the assessment objectives and the feasibility of gathering such information.

Preparing and pilot testing the assessment materials. Draft assessment booklets with accompanying answer blanks will need to be prepared in multiple forms containing the background questionnaire, test items, and administration instructions. Strict quality controls and test security controls should oversee the entire preparation process. A rigorous full-field pilot test conducted in a representative sub-sample of classrooms is needed to iron out difficulties in the examination booklets and administration procedures, and to provide guidance for preparing the final versions of the examination booklets.

Sampling vs. universe. A decision will need to be taken with regard to whether to test a sample or a universe of students. If the assessment is not also part of an effort at certification or selection, and is not designed to provide feedback to every teacher about his/her class, then a scientific sample of students would achieve the objective of improving learning at a fraction of the cost and effort of a full assessment of the student population. However, if the sampling is not done rigorously, then the assessment may provide inaccurate and confusing results. For an assessment based on a sample, a stratified cluster sample is normally most appropriate. Areas likely to be less responsive than average would need to be oversampled and private schools could be included for comparison purposes.

Scoring the assessment. Multiple choice items will need to be scored mechanically and essay items read and scored by professional readers backed up by supervisors and a reliability system. Items will need to be analyzed for quality and discriminability, initially on the basis of percentages and means, and later, depending on the testing agency's capacity, using more sophisticated techniques.

The nature of the testing agency. A decision will need to be taken on the location and nature of the testing agency. It is usually best for the central assessment agency in Government to be a very lean organization, with a few high quality staff, and which would contract out the major work of undertaking an assessment to a non-profit autonomous institution. If such an institution does not exist, then a long term government policy could be to establish and/or strengthen them. Usually an independent assessment policy board overseeing the entire assessment process is established, as well as subject councils and advisory committees, one for each subject area or academic discipline, to elucidate specific examination objectives within each area.

Disseminating test results. What information to report, to whom it should be reported, and how to report it is the first step in the design of a national assessment. Reports of results will need to be bold, clear, direct, concise, short, journalistic, balanced, graphically illustrated, non-technical, tailored to specific audiences, disseminated to the smallest level consistent with the sample, and presented by background characteristics. For maximum impact reports and results would need to be incorporated into pre- and in-service teacher training and other programs and into guidelines for teaching and supervision.

Tying assessments to research. Adequate funding and professional linkages will need to be built into the system to ensure that research on the effects of various educational inputs on learning is carried out using the assessment results.

Experience in Latin America

A review of the experience of Chile and Costa Rica is important, since these two
countries have implemented the most complete educational assessments in Latin America. The experience of Mexico is also of interest since it has a long standing Government agency which has undertaken a variety of assessments based on sample surveys. Through World Bank projects currently being prepared, all three of these countries are planning to strengthen their assessment capacity. Finally Colombia has a high quality university entrance examination system which in the future could be utilized for assessment purposes. The institution responsible for this program recently expanded its activities to support an entrance examination to secondary schools in the Bogotá metropolitan area.

Among the countries reviewed, Chile has had the most successful experience with assessments. Chile initiated its program in 1978. After initial problems, including resistance from teachers and students, the program has been in place since 1988, and has successfully measured learning for universes among fourth and eighth graders. The specific aim of the Chile program has been to use the results to affect educational policy and to strengthen teaching practice. The program included a strong dissemination effort involving civil servants, teachers, and parents. The program is continuing with strong support from the central government as well as from practitioners. The program has influenced government policy and there is some evidence that it has affected classroom practice. Specifically the results showing that private schools score higher than public schools have led to efforts to identify means of encouraging increased responsibility at the local and school level. The success of the Chile program appears to be a result of (a) the high quality of its staff, and (b) a strong focus on providing feedback directly to teachers, schools and districts and on informing the general public of the nature and role of assessment program. However, the program has not included research on causes of school failure and has relied almost exclusively on high cost censuses rather than on samples. Furthermore, the relationship between the Ministry of Education (MOE) and the Catholic University implementing the program up to now requires clarification.

Costa Rica initiated its program in 1986 and 1987 and assessed the universe of third, sixth and ninth graders. The program did not have a clear articulated goal, but its implicit goals included (a) using the assessments as a tool to argue in the public arena for additional funding for primary education, and (b) convincing the public of the need to re-introduce the use of partly standardized tests for certification of secondary school graduates. The assessment did not include the objective of using the assessment as a direct tool to improve classroom practice. Feedback to schools and teachers has been ad-hoc. Costa Rica has a limited human resource base, especially in Government, for this kind of activity. With the change of government, the program was in abeyance but will shortly be restarted with World Bank assistance.

The Government of Mexico has undertaken numerous assessments and evaluations over a twenty year period. It has also analyzed the national secondary school entrance examination for assessment purposes. The agency responsible for most of this work recently lost many of its good staff and has had an inadequate budget. It has been disseminating results, but on an excessively theoretical and general basis. It is now hoping to strengthen its staffing and is planning a stronger awareness raising program for parents, teachers and school authorities.

Colombia has an excellent autonomous agency undertaking university entrance examinations and examinations for entrance to public secondary schools in the Bogotá metropolitan area. Colombia’s exams are a model of modern computerized test preparation and scoring. The challenge for Colombia is to build on this capacity through utilizing selection examinations for purposes of assessment, similar to that of Kenya, as well as to develop a primary school level assessment system.

None of the four countries studied has adequately incorporated research into assessment
and testing. This means that much of the value of their effort is being lost because of a lack of additional complementary inputs. Furthermore there is an often unwarranted tendency to test the universe of students rather than to use less expensive samples. Using sample survey methodology will depend on good statistical expertise which is a scarce commodity. Furthermore these four countries have not put enough effort into dissemination of test results. Chile has the best record but could still do much more to have an impact on classroom behavior. Mexico’s dissemination efforts, while widespread, have been excessively general and theoretical and data on a per school or per district basis have not been systematically provided. Costa Rica’s dissemination efforts were undertaken on an ad-hoc basis by a testing agency outside Government.

Recommendations for Latin America

Coupling these four case studies with the general guidelines described above, Latin American countries seeking to establish or strengthen educational assessment systems will need to pay particularly close attention to the following elements:

- From the start Latin American educational authorities will need to define clearly the objectives of the assessment, including specifically how the assessment is expected to lead to improved classroom practice. They will also need to prepare a five-year implementation and cost plan, and act to ensure continuous and full support from the highest government authorities.

- From the start the managers of the assessment will need to design a full dissemination plan. The managers will need to identify the various clienteles who will use the assessment, hire expert writers and reporters for dissemination, and plan for in-service training programs of teachers based on the assessments. A strong effort will need to be made to educate all concerned parties as to the fact that an assessment is not designed to award or punish, only to assist, and that low scores in a region, city, or school, may not reflect on the teachers since there are many operating external factors.

- Latin American countries should seriously consider keeping the testing department outside government very small. In principle, most of the assessment design and implementation work should be contracted out to an independent, stable non-profit agency. If such an agency does not exist, then efforts should be made to build one up over time.

- Latin American countries will need to seriously consider using sample surveys rather than censuses, especially when the objective is to measure performance of groups of students, schools, and the system as whole. But samples had best not be undertaken if the sampling methodology is inadequate.

- Assessment programs will need to include funds for research, as well as agreed upon cooperative programs with local and/or foreign independent non-profit research institutions.

- Latin American governments will need to finance the training of psychometricians and statisticians. Contracting with non-profit agencies will help to ensure that this expertise remains available to Government, since it is difficult for Government to retain qualified staff with these scarce skills.

- Countries such as Colombia with well managed selection examinations will need to consider utilizing these examinations for assessment purposes.
Chapter I

INTRODUCTION

Issues in Primary Education in Latin America

Primary school enrollment ratios\(^1\) of over 90 percent in nearly all Latin American countries and over 100% in many of these countries in the 1980s suggest that most nations in the region have achieved the quantitative goal of having enough places in primary education to serve the school age population. Nonetheless, primary education in Latin America continues to be beset by high repetition rates, especially in first grade. Moreover, Latin America still fails to provide a complete primary education to its school age population. First grade repetition in the region as reported by UNESCO averages 20%. However new estimates and simulations now show that repetition rates have consistently been underestimated in official reports. They are now estimated at over 50% for Brazil and above 30% for Bolivia, Ecuador, and Peru.\(^2\)

Not only are repetition rates especially severe in grade one, they are also a significant problem even in the upper grades of primary schooling. What is more, although in absolute numbers illiteracy is decreasing in the region, the problem is far from solved, as 42.6 million (more than 15 percent of the population 15 years or older) still are illiterate.

The past decade has been a difficult one for education in Latin America. Public expenditure on education per inhabitant in Latin America increased in real terms between 1970 and 1980, but decreased significantly between 1980 and 1985, after which it began to recover some of the losses. As a percentage of GNP, public investment in education has also declined from 4.0% in 1976 to 3.7% in 1987. In most countries the relative priority given to education by Government has also gone down, with the share of education in the state budget declining from 18.9% in 1970 to 16.5% in 1975, 15.3% in 1980 and 16.2% in 1987. Highly vocal lobbying and interest groups have often succeeded in protecting the higher levels of education from the economic recession, to the detriment of the lower levels of education, especially the primary level. The available information on trends of education indicators is inadequate and it is difficult to quantify the impact on educational quality of the economic crisis in Latin America. However, anecdotal evidence suggests a real deterioration in the quality of education in the 1980s.

Numerous strategies for improving primary school quality are being considered, including increasing the availability of good textbooks and pedagogical materials; conducting in-service teacher education programs (including distance teacher education and programmed materials), with follow up activities, designed to raise teachers’ knowledge of the subject matter they teach and provide them with modern and effective instructional strategies; setting and maintaining standards for teachers as well as students regarding the amount of time each day, week, and year devoted to each subject; providing effective pre-school learning experiences; and increasing children’s learning capacity by providing them with appropriate school health services as well as snacks or modest breakfasts including vitamins, iodine, and iron supplements. Authorities are also trying to decentralize decision making and control as a means of encouraging local initiatives in quality improvement.

---

1/ Based on gross enrollment ratios for primary school age children, which is the ratio of the total number of students enrolled in primary school (regardless of their age) to the number of individuals in the primary school aged population.

2/ For a complete analysis of the repetition problem, see E. Schiefelbein, 1989.
In this paper, we present a different response to low educational quality in Latin America. This approach focuses on the establishment of national assessments of student learning as a vehicle for improving educational quality, particularly at the primary level. National assessment systems can help raise educational quality by collecting and presenting information on how well students are mastering the content and skills the nation's schools are supposed to teach them. One of the fundamental axioms of this approach is that to improve educational quality, it is necessary to start with a definition of quality that is based on student learning and cognitive skill development. Moreover, it is essential to obtain quantitative measures of educational quality on these terms.

Two countries, Chile and Costa Rica, have conducted national assessments for the primary level. Honduras, the Dominican Republic, Brazil, Jamaica, Colombia and Mexico have undertaken or are planning a variety of somewhat more limited assessments of student learning. These countries, as well as Belize, Ecuador, Northeast Brazil and the state of Sao Paulo, are considering establishing larger, more systematic assessment systems with the partial assistance of World Bank loans.

The Purpose of this Report

The main objective of this report is to inform education decision-makers in Latin America of the range and complexity of the issues involved in developing and implementing national educational assessment systems in this region. The report summarizes best practice in implementing and utilizing educational assessments, reviews the experience of several Latin American countries in implementing national assessment systems, and comes up with a set of concrete suggestions for the future of educational assessment systems for Latin American countries in general and for World Bank financed projects in particular. These suggestions are designed to help ensure that educational assessments can in fact influence policy makers and teachers to improve the quality of education. This report is timely because, as noted above many countries in Latin America and the Caribbean are currently embarking on programs to assess educational achievement in primary education. Given their economic problems and resulting constraints on public finances, these countries need to direct their resources toward cost effective investments, programs, and interventions in education.

The availability of even the highest quality educational assessment data cannot guarantee quality gains. This requires the commitment of national and local leaders as well as teachers, school principals, and the general public, to learn and be candid about assessment results, and to adjust education sector policies and influence practices based on assessment information and the results of analyses. This report therefore tries to provide some suggestions to help ensure, not only the effective implementation of educational assessments, but also their effective utilization to improve classroom learning.

The next chapter defines educational assessments and summarizes its potential uses. Chapter III reviews technical and managerial issues related to undertaking assessments. Chapter IV evaluates the experience in Chile, Costa Rica, Mexico, and Colombia in relation to these issues.
Chapter II

THE OBJECTIVES OF ASSESSING STUDENT LEARNING ACHIEVEMENT

Definitions

*National Educational Assessments.* This review focuses on the design and use of *national assessments of student learning.* Assessments measure the educational performance of a nation's or region's students and evaluates the progress of the schools, school districts, municipalities, or states in achieving curricular or other goals of the education system. Assessments can also be designed to identify problem areas in student achievement and recommend ways to improve student learning. Because the objective of assessment systems in general is to evaluate the performance of the system or components of the system relative to educational goals, it is only necessary to administer the assessment instrument to a scientific sample of the students rather than all of the student in the system.

*National Certification Examinations.* National assessment systems differ from national certification examination systems. The objective of national certification examinations is to identify students who have successfully completed a certain level of education or, alternatively, to select completers for entry into the next higher educational level (eg., from primary to secondary). Certification examinations given at the end of the primary cycle are common practice throughout Sub-Saharan Africa. To meet their objective, national examination systems must use the examination instrument to evaluate all students seeking a completion certificate or seeking selection into the next higher level of education, for which there may be a restricted number of openings.

Testing for selection occurs in many Latin American countries at the end of the secondary cycle, as a prerequisite for entrance to universities. These testing and selection examinations generally are organized in such a way so as to provide as little feedback as possible to students and teachers, beyond giving them a grade. This is in part because the designers of such tests are especially concerned with security and test administration. Generally, Latin America countries do not have national or regional tests at the primary level for purposes of certification or selection to secondary education. The exceptions include Mexico, the Anglophone Caribbean countries, and the Bogotá metropolitan area of Colombia (for entrance into public secondary schools in Bogotá). In most of Latin America individual teachers make judgments regarding whether children will pass or fail at all levels of primary education.

If tests are already being administered for purposes of certification or selection, it may be possible to redesign or restructure the instruments and the overall design in such a way as to serve some of the same objectives of an assessment system. In fact, the first step in considering whether to set up an assessment system in a country is to determine whether existing certification and selection tests can be adapted to serve assessment objectives. Piggy-backing on an existing testing program may be a cost-effective alternative to building a parallel assessment system, and should not be overlooked in any strategy for achieving national educational goals.

It is also important to differentiate between educational assessments, educational statistics, and educational research, all of which provide information of potential use to policy makers and practitioners. Typical educational statistics systems often cover such elements as enrollments by grade and age, numbers of student grade repetitions, estimates of student flow, teacher characteristics, physical facilities, equipment and materials, and educational costs. The advent of the personal computer means that
more data can be gathered within a shorter period of time, and fairly sophisticated analyses can be undertaken to guide policy makers. Latin American countries, as is the case with most developing countries, are only now beginning to go beyond the rudimentary phase of educational statistical analysis.

Educational research is defined as the rigorous testing of hypotheses about causal relationships. In particular, much educational research attempts to measure the determinants of student learning and achievement. State of the art educational research requires complex sampling designs, sophisticated analytical expertise, and as well as sophisticated computer modeling tools. Educational research, especially when undertaken on a longitudinal framework, can be quite expensive.

In principle the aim of educational assessments is not to support the testing of hypotheses, although such hypothesis testing can well be incorporated into them. In practice, as discussed below, the data developed by an assessment is so rich that significant benefits can be achieved if these data are linked with a comprehensive research program.

How Educational Assessments Can Be Used

As discussed above, the objective of educational assessments is to evaluate student learning, to assess learning achievement against established criteria, and ultimately to improve student learning by influencing the quality of schools, curriculum, and instruction. Obviously, by itself, measuring student learning does not yield increased achievement any more than weighing grain will yield increased agricultural output. It is, however, a necessary condition to establishing quantitative targets, assessing the tradeoffs of alternative allocation strategies, input combinations, and production technologies, and allocating resources and effort to achieve established targets. Improvements in educational quality may be achieved in a variety of ways, such as: through refinements in classroom practice and pedagogy; through improvements in curriculum design and instructional materials; through more effective teacher supervision; through an educationally reinforcing school environment; through increased use of appropriate educational materials; through better teacher preparation and skills upgrading programs; and through more cost-effective allocation of resources. Some of the specific uses of assessments to influence educational policy are discussed below.

To achieve any specific objective, of course, requires a strong dissemination effort. From the start, and before planning the assessment itself, it is necessary to decide what the assessment goals are and to develop a dissemination plan, and then later to include both funds and staffing to ensure that dissemination is undertaken. As will be seen, this fundamental element is often overlooked.

Using Assessments to Garner Public Support for Increased Efforts to Improve Educational Quality. An educational assessment system that measures student learning and reports on how well student learning matches clearly stated and understood performance standards not only helps the public understand what students know, it also allows them to track progress. Consequently, policy-makers can set clear targets, and the public can monitor progress toward these targets. Based on these measures, changes and adjustments in the direction of the system and in the financing of the system can also be justified.

If assessment learning targets are set too high, then the public may become disillusioned when reports indicate that very few students or schools (or only particular groups of student or schools) are achieving the targets. This political dimension suggests careful planning. Involving a broad spectrum of well respected educationalists and leaders in the goal setting exercises, as well as in the assessment policy
board, and eliciting the participation of the press throughout the process, will help assure the public that the national assessment system and the national educational goals are developed in a fair and open process. It also suggests that the audience(s) for the assessment reports should be clearly identified from the start, even before the assessment instruments are designed, to assure that the kinds of questions likely to be raised by each audience can be answered by the assessment.

**Using Assessments to Improve Instructional Design and Teacher Training.** Educational assessment systems can be used to provide information to educators that can help them improve instructional design and classroom pedagogy. Recently, a number of states in the USA have developed in their assessment system a number of finely tuned indicators designed to identify and accurately measure performance in a number of narrowly defined subject area domains. In this way, educators can tell if particular practices, teacher training approaches, or curricular materials are effective. Based on this approach, for example, a school system can redesign its in-service teacher training system in a way that the assessment analysis indicates is most effective.

**Using the Backwash Effects of an Assessment.** An indirect way in which assessments can affect teaching and learning is through their "backwash" effects. Backwash effects is the term used to describe the effects that an examination has on what teachers teach and what students make the effort to learn. "Teaching to the test" exemplifies backwash effects. Backwash effects are always present in a school system that relies on an examination to certify completion of one level of education or to determine selection into a subsequent level of schooling (i.e., secondary to post-secondary), or into a particular training program (nursing or mechanical training). In many instances, the backwash effects of a critically important selection examination reach back through many years of schooling, such as a university entrance examination that effects teaching from upper secondary down through primary school. If the examination is designed to measure the entire curricular domain, with a variety of academic and nonacademic applications, and higher order thinking skills, then it would make sense to capitalize on the backwash effects of the examination.

Assessments, as discussed above, are not usually combined with selection. Therefore, a sample based assessment system whose outcomes do not yield consequences for the students taking the assessment tests, may not produce any backwash effects *per se*. Nonetheless, the assessment system may be designed to encourage changes in the behavior of students or teachers similar to those yielded by backwash effects. First, if students, teachers, and parents take educational examinations seriously, then assessments may well exert powerful influence on what is taught and what is learned without backwash effects. By designing assessment instruments to examine the skills and knowledge areas that all students should acquire, rather than only those skills relevant for students likely to advance into the next level of the system, the schools will increasingly find themselves providing instruction that benefits all children. Second, assessments may be designed to provide rewards, of financial or other resources, or of prestige which may produce backwash-like effects (see page 6 below). Finally, countries may consider developing a long-range strategy to administer the assessment instruments to all students in all schools. In so doing, it may be possible to combine assessment and certification objectives in a single system, measuring the achievement of national educational targets while simultaneously assuring backwash effects.

A variation on the backwash effects of testing would be minimum competency testing.

---

3/ See "Managing the Assessment" in Chapter III, below.
The growth of the minimal competency movement in Alabama education provides a good example of this approach. State standards were set indicating the minimal level of performance on the Alabama assessment required of graduating secondary school students in each subject area. The minimal standards targets were not established to punish poorly performing students. Rather, they were designed to help district officers and school principals monitor the progress of Alabama students and to make sure that they achieve a reasonable standard of education. Research in Alabama shows that students at all ability levels, even advanced students, registered marked improvements in assessment performance.

**Diagnosing and Treating Learning Problems.** Another way in which assessment systems can be designed to improve educational practice is by providing content diagnosis and specific feedback from the assessment system to teachers, parents, students, and the community. By supplementing the feedback system with an analysis of the types of errors students make and with suggestions on techniques for improved teaching to help students avoid these types of errors, the educational assessment system can have a direct and positive influence on student learning.

This was done with the Kenya examination reform implemented in the late 1970s, with successful results (Somerset, 1987) -- a classic case of utilizing a selection system to meet assessment-type objectives. First the examinations were revised to assess a broader range of skills and higher order thinking skills. Then these new instruments replaced the older examinations which were mainly measures of students' rote recall. Subsequently, individual examination items were carefully analyzed to determine the types of skills in which Kenyan students were weak and appeared to consistently lack comprehension. For instance, by means of an analysis of the examination items, it was discovered that while students were measuring up well in the area of formal mathematics theory, they were not capable of applying their math skills to everyday situations. It was suggested that the curriculum and teachers were being too theoretical and were devoting too little time to practical applications. In one issue of the newsletter regularly distributed to the teachers as part of the examination reform program, teachers were advised of the problem, were given examples of the types of errors students were making, and were told of more effective means of teaching the necessary skills. A sample of the feedback mechanism used in Kenya is attached as Annex 1.4.

**Rewarding Good Performance.** Assessments can also be used to reward good performance or help those with bad performance to improve. Results from assessments have been used to hold regions, municipalities, districts, or even schools and school leaders (given the appropriate sample) accountable for performance in a number of states and districts in the USA. For example, in Boston and South Carolina, districts showing significant improvement in assessment scores (taking initial conditions and school social indicators into account) are rewarded with additional finances; those districts that continue to perform below a certain standard are sanctioned by transferring their authority to run their schools to an outside team of experts. Another example was Michigan's 1973 Chapter III plan (for educationally disadvantaged students). The plan was designed to initially allocate funds on the basis of schools with high numbers of poorly performing students, and then

---

4/ The Kenya feedback system has been criticized as helping students to answer examination questions better rather than improving learning. However, to the extent that examination questions measure fundamental higher level skills this is not a valid criticism.

5/ Student assessments cannot be used to hold teachers accountable for good teaching because it is not unlikely that differences in the quality of students taught by teachers being evaluated can be statistically controlled (Glass, 1974).
to allocate funds on the basis of student progress from baseline levels.

One of the problems associated with systems of rewards and punishments is that teachers, principals, and districts may have greater incentives to cheat—to record or report higher scores that those which were or would have been actually earned. They can justify cheating if they believe it will help them obtain resources for their students or if they believe that the system is unfair. Obviously, test security would have to be seriously considered when implementing this type of plan. Another problem associated with using accountability as a mechanism to raise school quality is that it would require all the nation's schools to be included in the sample (unless district-level accountability is sufficient). Moreover, the use of the assessment to assign rewards or to dole out punishments would lend the system a legitimacy for which legislators or other policymakers may not be initially prepared. These burdens suggest that funding decisions—certainly on a school by school basis—should not be a part of a new assessment system, and should only be carefully introduced under any circumstances.

Using National Assessments in Educational Research and Development

As noted above the primary aim of educational assessments is to measure progress in meeting the curriculum goals of the educational system—not to support the testing of research hypotheses. However, considering the richness of the outcome measures gathered by an assessment, the marginal effort needed to generate and test research hypotheses suggests that all assessments should include a research component.

It is true that certain questions can be answered very effectively by one-time surveys or even longitudinal surveys and often at a considerably lower cost than a permanent assessment system. Moreover, assessments do not and cannot effectively build in all of the statistical controls necessary to assure that good research can be conducted (especially as these controls are constantly being discovered or refined). In fact, many researchers have come to the conclusion that only full-fledged random assignment experiments can hope to determine causes of variation in learning achievement. Finally, it has been suggested that national assessments can be used to generate rather than test hypotheses regarding educational inputs and outcomes. However, this is an inadequate approach in a situation of limited expertise and ever-increasing financial pressures affecting schooling systems in Latin America. Policy-makers and educational practitioners need to identify the factors which can raise educational achievement, and which do so in the most cost-effective ways, in the simplest and most inexpensive manner, both financially and in the use of human resources.

National assessment efforts produce an immensely powerful set of outcome measures. To the extent that assessments also provide an array of control and background information on students and family characteristics, teacher characteristics, and classroom factors, along with information on costs, it will be possible for researchers to assess the effectiveness of alternative inputs into the schooling process, and identify cost-effective solutions to improving student performance. Moreover, a key component of such a research agenda is analysis that attempts to find links between educational achievement and labor market productivity or

---

6/ See "Designing an Effective Instrument" in Chapter III, below, for a discussion of test security issues.

7/ Examples of basic regression analysis from the Study of Diversified Secondary Education Curricula (DiSCuS) in Colombia, the Estudios Conjuntos Sobre Integración Económica Latinoamericana (ECIEL) in Peru, and a sample of a multi-level modeling procedure in Brazil are attached as Annex 2.
other social outcomes. This type of research would allow policymakers to assess the economic rates of return associated with different measures of educational skills acquisition. However, much of this type of analysis requires fairly advanced research capacity. A fundamental strategy in the building of an assessment system should therefore be the funding of a research capability that may include skilled and experienced subject matter specialists, educational analysts, statisticians, psychometricians, econometricians, sampling experts, and computer programmers. However, a country that lacks a state of the art educational research apparatus need not be constrained to start one without a sophisticated team of highly advanced researchers and research tools since adequate and useful research can be done without the use of the most sophisticated techniques. Nonetheless, the more complex the questions being asked of the data, the more advanced the research methodology required to answer them accurately, and the more sophisticated the research system will need to be.

In short, while assessment systems are not specifically designed as pure research instruments, they could be designed in such a way as to collect data that can be used by the research community to conduct basic research on education, to analyze the effectiveness of various educational inputs, and to estimate cost-effective ways of increasing student learning. Adequate funding could be built into the system to ensure that such research is carried out.
Chapter III

TECHNICAL AND MANAGERIAL ISSUES IN ASSESSMENT SYSTEMS

This section reviews issues and key decision points related to implementing educational assessment systems in developing countries. It includes discussions of what assessments should measure, the design of the instrument, ways of administrating assessments, and necessary institutional structures. It is assumed that the tests given are for purposes of assessment only and do not have any selection or certification objectives and that they cover primary and/or secondary education. This section is not meant to replace the standard textbooks on testing. Rather it is designed to alert education decision-makers to the complexity of the process and to the importance of rigor in planning, implementation, and analysis of results of assessments.

What and When to Measure

Defining Overall Objectives. A fundamental first issue is to decide what the objectives of the assessment are, as discussed in Chapter II, and then to decide what subjects and grades to test.

At the primary level normally the last year of the system and some key intermediate year (end of third grade) are usually selected. The minimum number of subjects is usually language and mathematics. Testing any additional grades and subjects will need to be carefully weighed against costs and technical capacity.

Achievement Testing versus Aptitude Testing. Aptitude tests are designed to measure the degree to which students have the capacity to acquire new understandings, gain academic skills, and apply the general academic principles one needs to succeed in schooling. Achievement tests, on the other hand, attempt to measure which understandings and academic skills were acquired, and how well academic principles are applied (Bejar, 1983).

Achievement tests have been shown to be better predictors of future educational performance than aptitude tests. Since achievement tests are keyed to the curriculum, strategies to improve performance in the subject area are more evident than with aptitude tests, which are designed to measure the ability of students to learn the subject. Also achievement tests can better incorporate the results of students’ hard work and discipline. Not to be unfair, a few points should be made in defense of aptitude tests. One is that aptitude tests, which should not be tied to what is actually taught in the classroom, limit the ability of teachers to "teach to the test." Moreover, by nature of their independence from the curriculum, they do not restrict teaching to follow traditional lines—opening up more possibilities for experimentation in the application of curricula and the utilization of instructional techniques. Another advantage of aptitude testing over achievement testing is that aptitude tests are better at identifying potentially high achieving, low-income students.

Criterion- versus Norm-Referenced Testing. One type of achievement examination is designed to compare student learning outcomes with average (i.e., typical) achievement levels. This is called norm-referenced interpretation.

---

8/ To the degree that backwash effects are being exploited, "teaching to the test" is a desired outcome and should count against aptitude tests.
For example, scores on a test given to students in a particular grade level can be compared to the "norm" score for students in that grade level. In a typical situation, a student obtains a score of 65 items correct out of 90 possible. The "raw score" of 65 is then interpreted by the use of a table that summarizes the performance of a group of other students who have previously taken the same test (the "norm" group). The student's score of 65 is equivalent to the 81st percentile, which means that the student equaled or exceeded the performance of 81 percent of the students in the norm group. The emphasis is on the relative position of students rather than on absolute mastery of content. Also, student scores can be rank-ordered for the purpose of awarding scholarships or making selection decisions for promotions when school places need to be rationed. Student assessment systems designed to compare students with one another in this way are called norm-referenced examination systems.

Criterion-referenced examination systems, on the other hand, assess whether particular or prescribed standards have been met. These types of assessments are commonly in use to certify individuals to perform particular tasks or functions--such as driving motor vehicles, operating machinery, or functioning in a foreign language. Criterion-referenced examination systems can also be used to test which knowledge or skills students actually learn. They can be used, for example, to assess whether students acquire or master specific skills in mathematics (e.g., long division, factoring polynomial expressions) or language arts (e.g., using the correct form of a verb). The chief drawback of criterion-referenced tests is that they are based on the assumption that the objectives of learning are to acquire a distinct set of abilities and skills that are few enough and important enough to be specified individually and mastered separately. Alternatively, learning may be considered too complex to structure in this concrete, but narrow way. In this broader view of learning, norm-referenced testing can more easily examine the nature of understanding. However this richer use of examinations is very difficult to implement.

Criterion-referenced achievement measures have an additional advantage for the development of a national assessment system--they may be interpreted in terms of attaining mastery of the subject area. More broadly, a criterion-referenced examination may be used to identify performance standards in education. To do this, the criterion-referenced test is assigned threshold levels of attainment. Those with scores below this threshold are considered as not having understood the material sufficiently to be considered proficient, or not having acquired the skill in question. Performance standards are also related to minimum competency levels--which indicate the minimal or basic level of skills and knowledge that children should acquire in a given level of schooling. Performance standards may be motivational as well. That is, they brings the idea of levels that the country's schools and youth can work to achieve.

In many countries there are already in place well-defined performance standards in the areas of health and physical fitness, and there are numerous types of performance standards in industry--such as for automobiles, medicine, and financial products. But there are few in education (performance standards are beginning to be used in the USA, as well as in the Philippines and in the UK). The kinds of questions that can be answered with a well structured system of performance standards are of the type, "How many 11-year-old children can calculate the amount of money earned from a savings investment at a given interest level?" or "How many children know the germ theory of disease?". It is true that test and measurement experts have not developed reliable ways of identifying where on a performance continuum a cut-off score should be set to establish performance standards. Nonetheless, it has been cogently argued that the advantages of establishing performance standards, even fallible ones, far outweigh the negative consequences.
associated with some degree of subjectivity (Popham, 1978).

To distinguish minimum competency from advanced proficiency, performance standards for any discipline could be set at various levels of competency. For instance, performance standards can be constructed to identify individuals who achieve a basic level, mastery level, or an advanced level of competency for each subject area or discipline. An additional measure, on a different scale than competency, could be a measure of exposure to the skill or subject matter. The first three levels of competency figuratively correspond to the British system's O level ("Ordinary"—similar to minimum competency), A level ("Advanced"). Exposure would indicate, for each subject, whether or not the skill or subject area was taught. The criterion for each level could be determined by experts in the subject or discipline.9

Criterion-referenced tests rather than norm-referenced tests are the preferred choice for assessments since criterion-referenced tests examine students' mastery of the system's educational objectives. The performance standards of criterion referenced examinations should be explicit and widely reported and understood.

"Objective" Format Tests versus Performance Format Tests. In testing and measurement circles throughout the world, academics disagree over the choice of examination formats to use in assessing student learning.10 Those who advocate the "objective test" formats, which include multiple-choice items, as well as matching items, question and short answer items, and rank ordering items, argue that this format is easier to administer, easier to score, more amenable to statistical analysis, more capable of being standardized, and more reliable than non-multiple choice formats. Supporters of "performance" tests, which range from open-ended essays or other writing performance tasks, to physical demonstrations of skills and competencies and portfolios of completed works, insist that the objective format, especially the multiple-choice format, is not "natural." That is, in life, as well as in higher levels of education, individuals are rarely required to solve problems by choosing from a small set of pre-selected alternatives or responding with brief, unsubstantiated answers. Moreover, they say, the non-multiple choice examination format—the essay examination for example—is better suited to assessing students' skills at analysis and synthesis.

Perhaps the most critical issue in comparing these two test formats is how accurately and reliably the examination measures achievement. To create an accurate and reliable multiple-choice test requires considerable skill and expertise. Careful design also permits the inclusion of analytic and synthetic assessment in a multiple-choice format. But the same relatively small number of experts could design a multiple-choice examination for millions of students. In contrast, the key to an accurate and reliable open-ended question is in the quality, training, and experience of the test scorers. The larger the number of tests, the larger the number of skilled graders are necessary. Hiring, training, and maintaining a cadre of test graders becomes a tremendous burden for a national testing system.11 Moreover, if there is a need to score examinations quickly (in order to provide results to the test-takers and feedback to the system), then the more feasible option is to use

9/ The types of institutions to be responsible for identifying examination objectives are discussed in the section on "institutional structures." These same institutions could also be responsible for identifying the competencies associated with each standard of performance.

10/ This section derived in part from Lockheed, 1991, page 3.

11/ But training of essay graders has been reported to be one of the most powerful forms of in-service training for teachers.
electronically scored multiple choice examinations.

Cost is a very important factor in determining which type of examination format should be used. First, objective examinations are much less expensive to administer and score than non-multiple choice examinations. The presence of relatively low-cost optical readers and computers reduces the cost even more. Further, the greater the number of individuals being tested, the lower the marginal costs of multiple-choice examination administration.

Items developed with the objective test format are more cost-effective, more easily standardized and made reliable, more amenable to statistical analysis, and easier and quicker to score than other test formats. However, if it is financially feasible and the human resources are available, then performance-type items, such as essay questions, in writing, language usage, mathematical analysis, and the natural and social sciences may also be included to ensure the teaching and learning of effective communication, original thinking, analysis, and synthesis.

Designing an Effective Instrument

This section covers the following steps in designing an effective assessment instrument: preparing the examination syllabus, training item writers, establishing background questions, preparing examination booklets, and field testing examinations.

Setting Detailed Test Objectives. The overriding purpose of education in any country is to produce changes in students' knowledge, skills, behaviors, values, and attitudes. To achieve these changes, educators formulate educational—or curricular—objectives. The way in which a curricular objective is stated indicates the complexity of the thinking process necessary for children to achieve that objective. Box 1, below, provides an overview of the different levels of thinking skills ranging from the lower order thinking skills of recall and comprehension, to the higher order thinking skills of synthesis and evaluation. The more clearly and specifically that educational objectives are stated, sequenced, and integrated, the easier it is to develop curriculum and instructional strategies from them, and the easier it is to assess whether the objectives are met in practice. Objectives stated in concrete, behavioral terms, called behavioral objectives, lend themselves particularly well to assessment.

To consolidate these curricular objectives, many countries produce what might be called a "teaching syllabus." This often consists of a set of manuals indicating what should be taught in each of the subject areas (science, mathematics, Spanish, etc.) and for each grade level (3rd year science; 5th year math; etc.) defined in the national curriculum. In some countries, the syllabus may only specify the learning objectives; in others, it will include specific curricula, textbooks to be used, the preferred instructional strategy, and even suggested lessons to be used by teachers. The national assessment system attempts to ascertain whether the objectives of the intended curriculum are achieved, and if not, why not. The assessment system is normally not involved in setting such objectives, although the results of assessments may well lead to curriculum change.

Given the breadth and scope of one or several years of schooling in any subject area, it is not be possible to examine all of the objectives laid out in the teaching syllabus. Instead, the assessment will normally examine whether students:

12/ Lockheed (1991) suggests that cost of producing and scoring an "objective" test may be, very roughly, 10 percent of the cost of producing and scoring a portfolio-type test (see Box No. 2 for broader implications of this cost difference).
<table>
<thead>
<tr>
<th>Question Category</th>
<th>Level of Thinking Required from Student</th>
<th>Examples</th>
</tr>
</thead>
</table>
| **1. Recall**     | Recall of facts, observations or definitions. | 1. What is the capital of Chile?  
| **2. Comprehension** | Giving descriptions, stating main ideas, comparing, and contrasting. | 1. Describe in your own words the theme of a story.  
| Describe, compare, contrast, explain, rephrase, translate. | 2. Compare life in urban and rural areas of the country.  
|                   | 3. Interpret a graph. | |
| **3. Application** | Applying appropriate rules and techniques to solve problems; recognize principles in new situation. | 1. Juan and Marta can do a piece of work in 15 days. They work together for six days. Then Marta leaves and Juan finishes the work in a further 30 days. In how many days can Juan do the piece of work alone? |
| Apply, solve, classify, select. | | |
| **4. Analysis**   | Making inferences, finding evidence to support conclusions. | 1. What is García Márquez' views of society in his novel *One Hundred Years of Solitude*?  
| Identifying causes or reasons, drawing conclusions, inferences deductions. | 2. Following your experiments, what is your conclusion about the factors affecting the growth of seedlings? |
| **5. Synthesis**  | Solving problems, making predictions. Producing original communications. | 1. What would happen if there was drought?  
| Predict, propose, plan, write. | 2. What actions could the government take to control population growth? |
| **6. Evaluation** | Judging the value of ideas, of a solution to a problem, of the merit of art and literature. | 1. Do you think that newspapers influence public opinion?  
| Judge, evaluate, decide, appraise. | 2. Should the government take action to control population growth? |

• acquire basic information in the key content areas of the disciplines (i.e., biology or chemistry, vocabulary or grammar in language, measurement or arithmetic) that a national assessment policy board determines should be covered;

• develop skills in the processes, cognitive domains, or thinking levels associated with the discipline (i.e., problem solving, estimating, inferring, or integrating--see Box No. 1); and

• apply these processes in a variety of contexts (i.e., in academic, personal, and daily-life situations).

To facilitate the development of examination objectives, a first step would be for subject matter specialists to prepare a matrix of the content and process characteristics of their disciplines, and assign the objectives to be assessed to the appropriate place in the matrix. Next, an examination syllabus, reflecting in a concrete but more focused way, the teaching syllabus, would be prepared. In particular, it would contain a selected subset of specific objectives from all of the relevant content, process, and context areas, sequenced in order of increasing complexity, that students would be expected to master. All this would require very close coordination with the country’s curriculum development department.

The examination syllabus, prepared by the national assessment team, is then used as the source for generating items for the assessment. The more detailed the examination syllabus, and the more that the objectives are stated in concrete and behavioral terms, the easier it will be to develop all of the examination items needed for the preparation of the assessment examinations.

The preparation of a detailed examination syllabus for each subject area and grade-level to be assessed is the first step in instrument development. Objectives listed in the examination syllabus should be distributed across the content, process, and context areas judged to be of fundamental importance. How much emphasis to place on recall of factual information and terminology, and comprehension of basic concepts and ideas, versus higher-order thinking skills, problem analysis, and the application of knowledge and skills is a matter that needs to be decided at the national policy level, not by the subject matter specialists or testing specialists.

Developing Test Items to Correspond with Test Objectives. The first consideration in the construction of an achievement testing system is to conceptualize and develop an item pool from which the examinations will eventually be composed. By following the logic and structure of the examination syllabus, and by developing sets of items that meet the criteria of each of the content, process, and context dimensions, the item bank would adequately represent the education system’s learning objectives. This step and the following ones form a complex process of constructing, sampling, equating, and pretesting examination items. These topics are more completely described in the educational testing literature (see, for example, Educational Evaluation, by J. Popham, 1975). What is described here are the fundamental issues that will need to be dealt with as well as a summary of the process as recommended by the educational testing literature.

The most common way in which items are developed is by hiring item writers. There are several requirements that item writers should meet. First, they must be specialists in their subject area and in the logic of the curriculum. Second, they must be experienced teachers. Third, they must prove to be sensitive to item bias factors to ensure that the items they write are not biased from the perspective of regional, ethnic, and gender, and socioeconomic differences. Fourth, they must have training or be experienced in item writing. This training or experience must be designed to develop skills in item phrasing, identifying appropriate distractors
(response options to multiple-choice questions that represent the types of errors students would typically make), and developing items that can measure the intended behaviors. Finally, subject-area committees (see "Managing the Assessment" below) oversee and provide feedback on test items developed by item writers for each subject area.

Test items need to be generated to correspond to the content and process areas identified in the examination syllabus. To help ensure objective, discriminating, unbiased, and curriculum-appropriate items for the assessment's item bank, highly qualified subject-matter specialists need to be given specific training in item-writing and their work overseen by subject-area committees.

**Developing Background Questions.** Developing a bank of curriculum-based items is not sufficient for an assessment program. To assess the schools' and teachers' success in providing basic education services, and provide information that would help schools and teachers improve student learning, it is equally important to measure student, classroom, school, and community inputs and costs. A sample of the school and teacher variables used in the ECIEL study is attached as Annex 3. Consequently, the assessment will also collect selected items of information from the following categories:

- Individual student-level measures, including student and family demographic and socio-economic characteristics, student school entry and grade repetition history, student self-reported health and nutrition knowledge, student distance from school, student access to texts and learning resources, student time spent on homework;

- teacher characteristics such as qualifications, experience, tenure, salary, in-service training;

- classroom level factors such as structure (single or multi-grade, departmental or multiple-subject programs), size (number of students), time allocated across subjects, teacher familiarity with the teaching syllabi, teachers use of lesson plans, frequency of supervisory visits, hours per week of assigned homework, availability of school time to complete homework; and

- administrative district level or municipal/state level resources and policies, such as per-pupil expenditure, school calendar and annual hours, requirements regarding choice of textbook or other pedagogical resources, requirements regarding supervision practices and other administrative mandates.

Another kind of background information could be student and teacher reports on their own attitudes, their self-esteem, and their aspirations.

It has often been pointed out that public officials are captivated by the idea of collecting information that might be used rather than will be used (Windham, Levin, and Bathory, 1990). With this in mind, care must be taken during the initial implementation phase of the national assessment to select only those background items that can be measured easily and accurately, and that could provide immediate benefits in terms of explaining differences in student, classroom, school, or district performance. Consequently, the initial list of background and control variables should be strictly limited. A sample questionnaire for the secondary and university entrance exams in Colombia with a minimum number of background questions is presented in Annex 4. Over time, the list of background items could be expanded in response to policy needs and the capacity of the assessment system to collect and analyze this additional information. One point that should always be considered in this process is that the value added by the additional information should be weighed against the additional costs associated with collection (including indirect cost factors such as the burden on respondents, and the direct costs of data collection).
During the initial phases of implementation, the number of questions designed to solicit information on student, teacher, school, and community characteristics should be strictly limited by the capacity of the system to collect and utilize such information when reporting and analyzing results. As the capacity of the system expands, and as new questions arise, additional background and control variables can be added to the assessment.

**Draft Assessment Booklet Preparation.** The next step in the process of instrument development is the preparation of the draft assessment booklets. These booklets normally contain two main sections. One section would contain the non-cognitive questionnaire items, discussed above. All students tested will receive the same questionnaire. The second section of the assessment is the cognitive assessment itself. This section will contain separate subsections of items from each of the subject areas to be tested. Testing experts will determine the number of items that need to be selected from each category in the item bank. Multiple forms of the examination booklet should be prepared to reduce test fraud problems and to facilitate reliable estimations. Test administration manuals should also be prepared to accompany the examination booklets (discussed below). Drafts of the assessment booklets should be reviewed and proofread by assessment subject area specialists, teachers, and copy editors to assure that there is no ambiguity or error. A security system and quality control procedures should also be developed and put in place to oversee the preparation, printing, packaging, and ultimately distribution of the examination booklets. Machine-readable answer blanks should be used for multiple-choice questions; these have to be prepared and tested at this time, as well.

The success of the assessment program depends critically upon whether students and proctors understand the directions to be followed when taking the examination. Normally two administrative manuals are prepared and field tested: a school coordinator's manual and a test administrator's manual.13

The school coordinator's manual describes the sequence of steps that need to be followed by the individual responsible for test administration at the school—from the time the assessment instruments arrive at the school until the moment they leave the school. This individual is normally assigned from headquarters rather than a school principal or a district inspector. Actual test administration instructions are contained in the test administrator's manual, discussed below.

The test administrator's manual is to be used by the individual responsible for administering the assessment. In most cases this will be the school coordinator, although in some cases (i.e., large schools) it will be an assistant to the school coordinator or another individual. The manuals are written in the form of a scripted play, complete with dialogue for the test administrator, expected responses from the test takers, the timing of each test section, suggestions for keeping test takers posted regarding time remaining, guidelines for proctoring, and a questionnaire to query the test administrator on the circumstances under which the assessment was administered, class size, seating and spacing, dynamics, etc.14

**Draft assessment booklets with accompanying answer blanks** will need to be prepared in multiple forms containing the background questionnaire, test items, and administration instructions. Strict quality controls and test security controls should oversee the entire preparation process. Test administration manuals will also need to be carefully prepared.

13/ This section is drawn from K. Ross, et al, 1990.

14/ This may help analysts who may be attempting to understand a pattern of outlier responses from a given school or classroom of students.
Pilot Testing and Finalizing the Draft Assessment Materials. Prior to finalizing and packaging of the examination booklets, one or more sets of field pilots of the materials should be conducted and carefully analyzed. During this pilot phase, the draft assessment booklets are administered to students in a variety of classrooms—not just to classrooms to which researchers have easy access—such as students from different ethnic groups (dominant groups, Indian populations, other minorities), from outlying rural areas, and from a variety of urban settings, to evaluate administrative procedures, test timing, item clarity, and item quality, among other issues. The items will be scored and basic data on item discriminability tabulated and analyzed by psychometricians or testing experts. A report summarizing problems and making recommendations for the final assessment booklet is then prepared. It should be stressed that both the administration manuals and the draft instruments need to be carefully pilot tested to ensure that the instructions are absolutely clear and that most encodings of test administration are anticipated and dealt with efficiently.

Pilot test results of the examination instrument are examined carefully by the research staff, item by item. Simple statistics (percent correct for each item, and distribution of percents on distractor choices) may be sufficient for the research staff to identify problem items, although more complex item analysis may be desired. Typical problems would include items that nearly all pilot test-takers got right or wrong, or on which particular groups of students scored very differently (girls versus boys, one ethnic group versus another, rural versus urban children). Problem items should be scrutinized and then revised or replaced.

Following these pilots, providing all outstanding problems are resolved, the examination and administration manuals are revised, finalized, and then published. Since an additional battery of pilot tests may be deemed necessary based upon the findings of the first pilot, the assessment implementation schedule should build in a sufficient amount of time for this activity.

A rigorous full-field pilot test should be conducted in a representative sub-sample of classrooms to iron out difficulties in the examination booklets and administration procedures, and to provide guidance for preparing the final versions of the examination booklets. Carelessness here can result in disaster for the entire assessment system.

Administering the National Assessment

This section identifies the following issues that need to be considered during the administration phase of a national assessment: whether the assessment should be given to a sample or to a population, the type of sampling, scheduling, administrating, scoring, and item analysis.

Sample-Based versus Census-Based Assessment. One of the most critical issues to be resolved prior to implementing the pilot assessment is whether to measure all students in the country for the grade levels to be assessed, or just a sample of students in these grades. Certainly if the assessment is designed also to be used in selection (i.e., which individuals qualify to enter particular institutions—such as universities or training institutes), then the entire population would need to be tested. Likewise, if the assessment is used to certify students (i.e., to assure that each student acquires the set of

---

15/ The literature on the theoretical and operational issues related to test item analysis and pretesting is vast, and ranges from the simple to the complex. The testing experts overseeing the item development process should be familiar with this literature and should have sufficient experience with the techniques of the trade to guide this process.
competencies required for graduation), then all students would need testing. In addition, all students would need to be tested if the assessment were to be used for individual diagnostic purposes. Then results for each student would have to be subjected to detailed analysis. This analysis would then be used to diagnose individual student competencies in a host of finely-tuned skill areas to inform the teachers and parents on competency areas requiring corrective measures.\textsuperscript{16}

However, if the objectives of the national assessment system are solely to monitor national education performance and to provide information to improve the quality of education, especially through determining how well curriculum objectives are being met and how resource allocations should be modified, then it is highly appropriate to use a well designed and executed sample, which can provide much of information gathered by a census at a fraction of the cost and effort. In addition, if well designed, samples can provide at least some feedback to teachers and to schools (but not to students) on their performance. The major caveat is that the sample should be a scientific stratified sample based upon an up-to-date sampling frame constructed for the target population and representative of all relevant and appropriate strata in that population. If this is not done the assessment may provide highly inaccurate and confusing results and had best not be done. Furthermore the public should be educated as to the nature of samples and the extent of their validity. To be sure that the realized sample is representative, an adequate level of administrator accountability should assure that substitutes are not made during survey time to ease the burden of the administrator.

If the assessment is not part of an effort at certification or selection, and is not designed to provide feedback to every teacher about his/her class, then a scientific sample of students, rather than the full population, would achieve the objectives of using an assessment system to boost educational quality (by allocating resources, setting targets, and improving practice, through feedback to educational leaders and practitioners) at a fraction of the cost and effort of administering the test to the entire student population.

\textbf{Determining the Sample.} One of the objectives of an educational assessment system is to improve policy-makers' understanding of how teachers, classroom factors, and schools influence educational quality and student learning. To facilitate an analysis of the effects of classrooms and schools on student achievement, a sample design that draws multiple classrooms from selected schools and includes all students in each of the classrooms is normally most appropriate.

Specifically, the schools in the sample frame must be sorted into clusters according to geographic region, municipalities within region, and size within municipality. This approach is called stratified cluster sampling. The schools within each cluster should be selected with a probability proportional to a measure of size (such as the estimated number of enrollees in the school, average class size, etc.). Before selection, it may be necessary to exclude very small schools or schools in remote areas where the costs of testing would be excessive.\textsuperscript{17} It may also be necessary to apply a higher sampling rate to clusters of small schools, schools in rural areas, and schools in areas outside the mainstream. Among other benefits,\textsuperscript{16} This approach is being prepared for implementation in several states in the USA at a high cost (although many local school districts have developed their own testing materials for this purpose).

\textsuperscript{17} In some countries, data from these types of schools are critical for educational policy. In such cases it will be necessary to bear the expense and sample from them.
this over-sampling will help analysts deal with higher than average non response rates. In the end, a realistic sample size based on cost and other methodological considerations will have to be determined.10

Since the marginal cost will be low and the marginal benefits high in terms of research possibilities, it is usually appropriate to measure all the students from the grade-level to be assessed in each of the sampled schools. This sampling technique will allow researchers to study the relationships between teachers and students. The assessment survey can also collect additional information from the teachers responsible for educating the pupils in these classrooms in order to permit additional analysis. Ultimately, this type of information may help researchers provide feedback to teachers and teacher training institutions regarding effective or ineffective teaching practices. The alternative, to select a random sample of students from within classes, will result in the giving of fewer tests and therefore some reduction of costs, but more importantly will run the risk of sample bias (i.e., the teachers may select the best students), add to costs, and reduce the capacity of the assessment to measure teacher and class composition variables. Moreover, from the point of view of test administration, it is probably easier to test an entire classroom of students than to select a small sample of students from the classroom and find an appropriate and secure environment for them to take the test.

For an assessment based on a sample, a stratified cluster sample should be drawn. Usually the best approach is to include in the final sample all students from the grade level(s) to be assessed in the sampled schools. Areas likely to be less responsive than average are over-sampled and it is often appropriate to include private schools for comparison purposes.

Administering the Assessment Booklets. Before an assessment system can function, the assessment has to obtain the full cooperation of local school administrators, who should be made aware of the potential value that the assessment will provide to educators. Then enumerators and supervisors going from school to school need to be hired and trained. Finally, during a controlled period of time, the assessment booklets are administered nationwide.

The test administration team should obtain the full cooperation of all school administrators, hire and train enumerators, assure that test booklet packages are properly prepared, sealed, and labeled, and then administer the test nationwide during a controlled period of time.

Scoring and Analyzing Test Items. There are two types of items that need to be scored: multiple choice items and essays (if included). The multiple choice answer blanks are processed in the standard way by optical scanners. The essay items are scored by professionally trained readers, and the readers’ scores for the students’ essays entered on answer blanks. For the essay sections, professional scorers will have to be trained in reliable methods of reading and scoring the essays. Tight supervision and quality control procedures should be in place for this scoring, including a system to measure and control inter-reader reliability (i.e., every fourth essay is read blindly by a second reader).

Other sampling issues should also be considered if it is important to test a wide range of curricular domains and a broad array of competencies within each domain. A variety of fairly advanced sampling techniques have been developed that allow for significantly greater breadth in the curriculum areas being assessed without excessive test burden on the respondents. For example, the US National Assessment of Educational Progress uses a "spiral bib" sample design that covers a broad range of curricular areas by testing a relatively small pool of respondents taking short tests.
Once the tests are scored, the data is structured into a database. The scores are checked against acceptable ranges and inconsistencies resolved according to predetermined procedures. A codebook describing the administration of the test, documenting the database, and analyzing the tests and survey questionnaires (see next section) is also prepared.

Analyses of the item responses can be conducted in a series of stages to allow host-country expertise to develop at its own rate and, ultimately, to improve the quality, accuracy, and utility of the instruments. A range of empirical item analysis and item improvement techniques has been developed over the years for both norm- and criterion-referenced tests. These range from the simple (percent correct analysis), to the complex (IRT scaling or scale anchoring). For a technical consultation of this and other topics related to educational measurement, see Berk, 1984; Carey, 1988; Popham, 1990; and Suen, 1990.

Multiple choice items should be scored mechanically and essay items should be read and scored by professional readers backed up by supervisors and a reliability system. The scores are structured into a database and a codebook prepared to accompany the database. Items then need to be analyzed for quality and discriminability. Initially, percentages and means could be assessed. More sophisticated item analysis can also be introduced, such as IRT analysis or anchor scaling, depending on the testing agency’s interest, level of sophistication, and resources.

Disseminating Test Results and Providing Feedback

Analysis and dissemination of assessment results are fundamental elements of an assessment system. Sufficient resources should be earmarked to support these efforts. The key issues and decision points are related to: the planning of what information to report and to whom; the timing, content, and breakdown of results; pedagogical analysis; provision of research information; and assessment of the assessment itself.

**Planning the Reporting of Information.** To ensure that the key goal of the assessment is met—to raise student learning by promoting improvements in school quality, it is essential that the right kind of information is reported to the appropriate decision makers, and that this information is reported in the clearest, most useful way. To disseminate assessment results effectively, it is critical that the reporting of assessment information receives as much advance planning and effort as test construction itself. The rule for good reporting (Shepard, 1980) is to decide what should be reported before planning the assessment itself.

Deciding which information to report, to whom it should be reported, and how to report it is the first step in the design of a national assessment, and should guide the development and operation of the entire assessment system.

**Different Reports for Decision-makers at Different Levels.** There are many different types of decision-makers influencing basic education policy and practice in any country—students, parents, teachers, head teachers and principals, district or municipal officials, national education officials; district and regional education officials; regional or state education officials, national education officials, other politicians, and the general public. Since each type of decision-maker has different needs and interests, different assessment reports must be designed for each group. For example, national officials are mainly interested in three types of questions: What are the differences in performance in each major subject area between districts (after accounting for social factors that might influence these outcomes)? What progress is the nation making towards improving these performance measures? and What can be done to improve
performance levels overall and to reduce performance differences between regions and districts? Regional and district officers may ask the same questions, but would be more concerned with differences among schools or groups of schools. The public may be interested in how well their country's school children are performing relative to students in neighboring countries. The important point is that different reports should be prepared for each of these different groups, with information content and format keyed to the interests and needs of each.

Some elements to consider in preparing these reports include (adapted from Shepard, 1980):

- employ a journalistic, not scholarly writing style;
- state the key findings first;
- do not include irrelevant information;
- make reports short, concise, clear;
- use explanatory graphics to highlight important points;
- reserve technical explanations for technical appendices;
- do not hedge findings with conservative statements about statistical inference;
- do not obscure information;
- make comparative and interpretive statements.

To elicit and maintain interest in and support for the national assessment, it is essential to rapidly process the test scores and disseminate results to the general public, who are the ultimate "users", as quickly as possible. Considering all of the steps involved in this process, this is no simple task. The first step in the dissemination process is to provide a preliminary report on national scores. While this may be done without resorting to advanced analytic techniques, it is still preferable to analyze and scale the results carefully prior to dissemination.

To make test information more meaningful—and more useful—to the general public, as well as to teachers and students, the results are best reported as specific proficiencies compared to established performance standards. This helps answer the questions, "What do our school children know and what can they do?" and "How does this compare with what school children are expected to know and should be able to do?"

These results can be compared by student background characteristics, such as gender, rural/urban location, or family socioeconomic level; or by teacher characteristics, such as years of experience, educational attainment, type of training, and recency of in-service training. This would help users understand the results. The report should also be sure to educate the public as to the nature of possible misinterpretations of any report of this sort. For instance, the fact that rural areas or poor urban slums will have the lowest scores should not be considered as an indictment of teachers or schools. Rather, the handicaps under which these students operate should be pointed out and the focus should always be on the positive, i.e., what can be done and how can the system move forward. In fact, if not handled properly, a highly negative report without appropriate positive exhortations could kill the public appetite for this kind of exercise.

Either along with national results, or as soon as possible afterwards, municipality by municipality, district by district, or school by school comparisons (depending on sample sizes) of overall scores and aggregated proficiency scores should be published. These comparisons should not be made without simultaneously reporting average socioeconomic statistics, per-pupil expenditure statistics, age-cohort
enrollment rates, and/or other statistics for the states or districts that may affect average performance for a given municipality or district.

Test results need to be processed and disseminated as quickly as possible. The results generally should be presented by background characteristics and some broad conclusions should be stated. There should be a careful balancing of the presentation in order not to be overly negative. Results should be disseminated by the smallest level consistent with the sample, combined with basic indicators for each administrative area that might contribute to differences reported. Test reports should be tailored for different audiences. All reports should be bold, clear, direct, concise (and short), journalistic, graphically illustrated, and nontechnical.

Conducting Pedagogical Analyses of Results. One of the more valuable uses of the assessment is to allow educators to better understand what students are learning, what types or patterns of errors they are making, why these errors are being made and, importantly, what teachers can do to help students overcome these errors and misunderstandings. Consequently, the reports issued to the public would be transformed from statistical reports on the average score in a municipality or the number of students who attain a given performance standard to a pedagogical report to teachers on how to help students better understand a set of concepts. The form of these reports could be a guide, a newsletter, or materials incorporated into inservice teacher education programs. In fact, incorporating the assessment results into preservice and in-service training programs, especially for teachers, school principals, and supervisors, can help ensure full utilization of test results. A curriculum can be developed whereby these professionals learn how to interpret test results and to utilize them to improve instruction and educational management.

Analyzing the test results pedagogically can provide useful insights to educators regarding recurrent problems, errors, or misunderstandings made by students. These insights can be transformed into advice to teachers in how to raise student learning. They can also be incorporated into teacher and other inservice and pre-service training programs.

Research and Cost-Effectiveness Analysis. As noted in Chapter II, one of the direct outcomes of the national assessment effort is the production of an immensely powerful set of outcome measures. As a richer array of control and background factors become part of the national assessment, along with information on factor costs, it will be possible for researchers to assess the effectiveness of alternative inputs into the schooling process, and identify cost-effective solutions to improving student performance. Moreover, a key component of such a research agenda is to link educational outcomes with labor market productivity and other social outcomes. This would allow policymakers to assess the economic rates of return associated with different measures of educational skills acquisition. These analytic activities require quite sophisticated research capacity. One component of the strategy to build the assessment system should therefore be the funding of a research capability that would include skilled and experienced researchers, education analysts, psychometricians, and econometricians.

While assessment systems are not specifically designed as pure research instruments, they can be designed in such a way as to collect data that can be used by the research community to conduct basic research on education, to analyze the effectiveness of various educational inputs, and to estimate cost-effective ways of increasing student learning. Adequate funding will need to be built into the system to ensure that such research is carried out.
Self-feedback to Improve the Assessment System Itself. Once it is initially established, the national assessment system will need to grow and evolve as staff gain expertise, and as comments and recommendations from practitioners and the public filter in. Assessment staff should also be involved in this review process to improve future assessment activities. Several types of feedback will probably be offered.

One area would be how well the system objectives were assessed by the instrument. Were important objectives left out? Were the objectives assessed adequately measured? Were the objectives measured even appropriate to the grade level assessed?

A second area is the design, administration, and scoring of the instrument. Were the enumerators sufficiently trained? Were there any problem for students in using the assessment? Was the layout and design of the assessment suitable? Was there an excessive number of irregularities in the administration? Was the timing of the examination segments appropriate? Were there any serious security breaches? Was the professional scoring of the essay sections reliable?

A third area is that of analysis and dissemination. Were the assessment staff sufficiently trained to analyze the items? Were the results published quickly enough (i.e., on schedule)? Were there appropriate reactions from the public regarding general findings and the disaggregation by municipality? Was the pedagogical analysis undertaken and effective? Were teachers able to benefit from the general and pedagogical reporting and analysis? Did the assessment achieve the desired results in terms of raising interest in achievement and quality? Is there any evidence of changes in curriculum, in teacher behavior, in student achievement as a result of the assessment?

One way to ensure feedback is to have an outside team of evaluators make recommendations on how to improve the assessment for the next round.

The best assessment systems include self-monitoring as well as evaluation, often by outside teams, so as to improve the system in response to the needs and concerns of practitioners and the public.

Managing the Assessment

The Role of a Central Assessment Agency. In many Central and South American countries, the Ministry of Education (MOE) houses an office of evaluation or of testing. In other countries, an assessment office will need to be established in the Ministry or one of its statutory institutions, such as a national curriculum development center, or the government’s national center for education research. The overall responsibility for the assessment normally resides with this central education evaluation agency which liaises with the Minister of Education to assure sufficient finances and continued political support for the ongoing assessment effort, and is responsible for deciding on the fundamental objectives of the assessment system and for ensuring cooperation by districts, schools, and teachers. It is best that there be a very few senior staff in this office or agency who are well trained and highly regarded professionals who are paid competitive salaries.

In principle the remainder of the responsibilities for running the assessment should be contracted out to an outside organization (a research university, non-profit agency, or consulting firm) with a staff of research experts in testing and measurement, sample design, and education analysis. These expert staff will be responsible for implementing the assessment and assuring that the assessment is coherent, feasible, and scientifically correct. The devolution of responsibility for implementing the assessment outside the Ministry insures that the highest quality personnel in scarce areas such as psychometrics and sampling areas are incorporated in the
process. It helps ensure that results will be objectively and scientifically reported. There is a tendency in some countries to start with outside contracts and then to incorporate the whole system into the ministry structure. This may well eventually result in the establishment of a large, unwieldy bureaucracy with inadequate high level expertise. In cases where such outside agencies may not exist or may be very weak, it would be important to provide long term financial and other assistance to strengthen such an institution.

In principle, it is best for the central assessment agency with Government to be a very lean organization with a few high quality staff. The major work of undertaking an assessment would be contracted out to a non-profit autonomous agency. If such an agency does not exist, then a long term government policy could be to establish and/or strengthen existing agencies.

The Role of Assessment Committees and Councils. A successful assessment system depends upon political and public acceptance of the goals and strategies of the assessment and its implementation. This acceptance in turn depends upon the collaboration of many interests. The integrity of educational data, for example, can be assured at a minimum, or improved, when the government-sponsored institutions responsible for developing and maintaining the instruments and statistics collaborate with the research community in data collection, in the construction of the instruments, and in the interpretation of the findings. In addition, the quality of the assessment statistics will be raised substantially when the host institution can rely upon expert staff who are well respected by the academic community and who are somewhat removed from the demands and needs of politics. Finally, when the research community accepts and uses educational achievement outcome measures, their acceptance and use by policy makers is further assured.

To ensure that these diverse needs are met, it is appropriate to establish an assessment policy board to oversee this entire assessment process and help assure that these different needs are met. Such a board would be independent of the political process, but responsive to national policy needs.

An assessment policy board could be composed of 4 to 8 individuals selected by the MOE's central education testing agency. The presiding officer could be the director of the MOE agency. The Board would also include the director of the non-profit agency responsible for implementation of the assessment. The other board members would comprise individuals from diverse backgrounds with different skills and constituencies, including primary school educators, secondary school teachers, university faculty, and education research experts. To garner greater support from the public, it would be beneficial if it included leaders from professional associations, such as the academy of sciences (if such an organization exists), as well as from business.

The Board's primary purpose would be to decide what grade levels to examine, which subject areas to assess, when to conduct the assessment, and to oversee the development of the examination objectives and performance standards drafted by the subject councils (discussed below). Further, the policy board would ensure that the assessment system is fairly administered and efficiently run.

Each of the subject areas included in the assessment would need a separate subject council made up of subject specialists—including, as appropriate, primary teachers, secondary teachers, university subject experts, and, when appropriate, individuals working in the occupational areas related to that subject area (such as engineers for science or statisticians for

---

19/ This section is drawn from Horn and Winter, 1991.
mathematics). The councils would need to liaise closely with the curriculum development department of the Ministry of Education as well as with curriculum development experts in the universities and teacher training schools. The subject councils would concern themselves with what students are expected to learn in that subject area, and with competency criteria, that is, what students should know to be considered competent in that subject area. The subject councils would prepare tables indicating each of the critical processes and concepts students in the given subject area and in a particular grade-level should know, followed by an examination syllabus for each subject area reflecting, in a concrete way, the teaching syllabus, and used as a source for generating items for the assessment.

The subject council and research staff will need to work together on developing an item bank including conducting or overseeing a pilot test of the items, scoring of the items, and making the final determinations regarding which items to use. The subject councils would also explicitly identify important concepts or skills that are not initially covered in the examination syllabus (or examination objectives). The councils would engage in a long term iterative process working with specialists, practitioners and outside experts to refine what is taught and what is being measured in primary schools in the nation.

To ensure that the diverse needs of the assessment are met, it is usually appropriate to set up an independent assessment board with representatives from broad concerned sectors of society. Also, subject councils are needed to ensure consistency between assessment objectives and specific test items.

Expanding the Scope of a National Assessment System

During the initial implementation of a national assessment system, all efforts should be directed toward achieving the basal or essential aspects of the system: measuring, tabulating, and reporting valid and reliable measures of student achievement. At this phase, the management and technical burden should be kept at a minimum, and expectations should remain fixed on these accomplishments. One way to assure that expectations are kept simple and achievable is to limit the assessment to one grade-level and one or two subjects, such as Spanish and mathematics. Once this phase of the assessment is achieved, and public acceptance is gained, the assessment may be expanded, broadened, and enriched.

Expanding the Coverage of the Assessment. One option would be to expand the number of subjects tested and perhaps introduce a second or additional grade-levels to be assessed. If desired and feasible, the sample can also be expanded physically to be able to assess achievement in all of the districts in the country. A final expansion could be to provide school-by-school comparisons that summarize basic educational outcomes in all of the key (assessed) subject areas, teacher qualifications, and availability of school resources. These reports can motivate site administrators to improve supervision, institution in-service training programs, or lobby for additional resources from parent groups.

Using the Assessment as a Stronger Vehicle for Educational Research. To facilitate educational effectiveness analysis and cost-effectiveness analysis, the background information questionnaire can be broadened to become more useful for effectiveness and cost-effectiveness research. The reason for measuring the flow of classroom inputs, resources, and processes is to allow host country social scientists to examine relationships between these factors and the process of student learning. Consequently, it is important for these inputs to be comprehensively and accurately measured. Particular attention should be given to inputs and resource measures that are relatively easy to manipulate by policy makers, administrators, or teachers. Classroom-level variables, with information collected from the teacher or
administrator, might include the following measures: class size; teacher qualifications; teacher experience; teacher tenure at the school; teacher salary; recency, type, characteristics of, and duration of most recent in-service training; special teacher retraining programs; teacher score on the student tests, or teacher's report on whether he or she will or did teach to the identified curricular objectives; type of classroom (single or multi-grade); reading/math textbook used; publication date of textbooks; ownership of textbooks; and price of textbook to student (current year).

In addition to the school resource measures listed above, it might be possible to obtain careful qualitative descriptions of classroom processes and school policies in a number of areas, such as: method of deciding which students are promoted to the next grade; hours last/typical week that students are engaged in math and in reading activities; hours per week homework is assigned in math and in reading; teacher and principal assessment of disciplinary environment; class time taken up by interruptions; number of students on sick call on a typical day (and the time they are away from the classroom as a result); hours per week the school is open; number of years school has been in continuous operation; length of the school year; required allocation and quantity of time on reading, mathematics, and other academic activities; number of teachers/classrooms in the school; and availability of school time to complete homework.

The assessment could also collect information on the students' personal and family backgrounds, including health, nutrition and non-school learning experiences. Availability of this type of information will assist analysts in uncovering the relationship between school processes or resources and educational quality. In particular, while parental background appears to have substantially smaller effects on learning in underdeveloped countries than in advanced countries, failing to control for these influences would nevertheless bias studies of school effectiveness. Disease and nutritional deficits are important contributors to the low levels of academic achievement of youth in many developing countries (Pollit, 1990). Consequently, the assessment could include some measures indicating nutritional knowledge as well. Individual student-level variables might include: student's frequency of grade repetition (which grades, how many times; student changes of school; student attitudes about school and about particular subjects; amount of time doing homework; student obtaining help on homework; whether the student had the same teacher the previous year; education level and occupation of parents; availability of radio, TV, books and newspapers in the home; public availability of a TV or movie theater in the village; age and years spent in school prior to this one family structure, including the number of siblings; student self-reported health and nutrition knowledge; and student sex, age, height, and weight.

Additional information may also be collected at the administrative district-level (i.e., the municipality). This might include: expenditure per pupil; frequency of supervisor visits to schools and characteristics of such visits; and academic calendar policy.

Data on Factor Costs. The assessment could also attempt to obtain data on the costs of educational inputs, such as teacher salaries, family's educational fees and expenses, school books and materials, and in-service training programs. In many cases, it may be possible to reconstruct cost estimates with the help of children, teachers, principals, or experienced local education experts, depending on who can provide the types, numbers, and costs of fees, materials, transportation, textbooks, workbooks, and teacher manuals that are identified in the assessment. Teachers can supply information on the type, frequency and recency of attendance, distance, and costs of in-service training programs, inspector supervision visits, etc. The research team should develop a strategy to obtain this information.
Enriching the Assessment With a Longitudinal Component. Adding a longitudinal component to the assessment process will permit researchers to measure the relative impact of school quality investments during the course of a particular block of time, such as from the start to end of the school year for the same students, or across several years for the same students. The natural way to assess the impact of pedagogical resources and processes on learning is to focus on the growth in achievement that takes place during a given time period. For example, researchers may be interested in measuring factors that affect the gain in knowledge and skills that takes place during the final year of primary school. In particular, they may be seeking to determine which resources and classroom practices produce the greatest gains in student achievement. Consequently, the most reasonable way to measure gains in pupil knowledge is to subtract the beginning-of-the-year score on a criterion-referenced test from the end-of-the-year score. The assessment can add multi-point measurements to the system’s overall design. However longitudinal studies are very expensive and require a high level of expertise, especially in drawing a good sample.

After it is fully established, the national assessment system can be enriched by expanding coverage to more grades or subjects. The assessment can also become a stronger vehicle for research by increasing the amount of classroom, process, student background, and district-level information gathered, as well as by introducing cost information and a longitudinal component. All such enrichment will require adequate financial and human resources.

Costs

Box No. 2 provides very rough estimates of the relative costs of sampling compared to universes and "objective" compared to performance tests. Depending on the modality selected and the size of the universe, the costs for, say, a primary level test in Spanish and mathematics could range from as little as US$100,000 to as much as US$40 million. The least expensive system is one with "objective" tests directed at a small scientifically selected sample. The most expensive system is a "performance" based system directed at the universe of students.

With few exceptions, the costs of undertaking even a large scale assessment are always a very small fraction (less than 1%) of the total bill for the educational system, of which most of the costs are for salaries. However, these costs are discretionary, with uncertain and sometimes politically sensitive results.

It is quite important to prepare realistic cost estimates from the start for a five to ten year program and to try to get a commitment from central government authorities to adequately finance the assessment system over that period of time. This will also influence decisions on the number of subjects and levels to be tested. Inadequate financing may result in a one-shot assessment poorly integrated into the education system and therefore with little impact on learning.

---

20/ For a review of the issues concerning the measurement of learning individual change over time, see Willett, 1988.
Box No. 2 - Comparison of Costs of Administering Types of Assessment Systems 1/

<table>
<thead>
<tr>
<th>Type of test</th>
<th>Fixed cost for test preparation</th>
<th>Estimated cost per test</th>
<th>Number of students in education system</th>
<th>Costs for sample of 2,000 students</th>
<th>Costs for universe of all students</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Objective&quot; test (student shows knowledge and skills by answering multiple choice and open ended questions)</td>
<td>$100,000</td>
<td>US$1 per test-taker</td>
<td>400,000</td>
<td>$102,000</td>
<td>$500,000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2 million</td>
<td>$102,000</td>
<td>$2.1 million</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4 million</td>
<td>$102,000</td>
<td>$4.1 million</td>
</tr>
<tr>
<td>Performance test (student shows knowledge and skills through essays, demonstrations, laboratory exercises)</td>
<td>$50,000</td>
<td>US$10 per test-taker</td>
<td>400,000</td>
<td>$70,000</td>
<td>$4.05 million</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2 million</td>
<td>$70,000</td>
<td>$20.05 million</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4 million</td>
<td>$70,000</td>
<td>$40.05 million</td>
</tr>
</tbody>
</table>

Source: Derived from Lockheed, 1991, Table 2, and from Esquivel.

1/ Costs are based on rough estimates in countries such as Costa Rica and Ecuador for a primary-level test in Spanish and mathematics. Sample size will vary depending on the objective of the assessment, including how many sub-groups (e.g., urban/rural/public/private, by region) are to be studied. A sample size of 2000 will usually be adequate in terms of giving reliable estimates of scores if only a few sub-groups are to be studied.
Chapter IV

ASSESSMENT SYSTEMS IN LAC: STATE OF THE ART

As noted in Chapter I, Chile and Costa Rica have assessed educational attainment at the primary level at a national level. In addition to Chile and Costa Rica, Honduras, Brazil, Jamaica, Colombia, and Mexico have undertaken a variety of limited assessments of student learning. These countries, as well as the Dominican Republic, Belize, Ecuador, Northeast Brazil, and the state of Sao Paulo are considering establishing larger, more systematic assessment systems with the partial assistance of World Bank loans. This section summarizes the experience of Chile and Costa Rica, which have implemented the most complete educational assessments in Latin America, in relation to the technical and managerial issues described in Chapter III. This section also summarizes the experience of Mexico, since it has a long standing Government agency which has undertaken a variety of assessments based on sample surveys, with mixed results, and which is currently facing major problems of financing and personnel. Finally the experience of Colombia is also reviewed. Colombia has a very strong and well organized university entrance examination system. The institution responsible for this program has expanded its activities to support an entrance examination to secondary schools in the Bogota metropolitan area. The example of Colombia is included, first, because of the high quality of its testing agency, and secondly, because it is a good example of a selection system which in the future could be utilized for assessment purposes.

Chile—History and Background

In 1978 the Ministry of Education (MOE) commissioned a team from the Universidad Católica to develop a system to assess the quality of the education focusing on factors that influenced educational achievement. In 1980 MOE conducted pilot studies in two regions, and the Programa de Evaluación del Rendimiento Escolar (PER) started in 1982. In 1984, a series of tests were administered nationwide covering 390,000 fourth and eighth grade students in 3,200 schools. The main objective of the tests was to obtain information on cognitive achievement (reading, mathematics, and social and natural science) and to inform parents and supervisors about achievement levels on a school by school basis. However the work was not successfully completed because of inadequate administration and resistance from teachers and supervisors because of threats that teachers or schools would be punished for poor results on the tests. Shortly after, the Centro de Perfeccionamiento, Experimentación e Investigaciones Pedagógicas (CPEIP), a MOE dependency, developed the Sistema de Evaluación de la Calidad de la Educación (SECE), with the objective of identifying factors that could be used to improve the quality of the education in preschool, primary and secondary education levels. A weak infrastructure and limited resources explained why SECE also had a very short life. Only one test application was administered among 9,300 eight grade students in 300 schools in 1986.

Based on these experiences MOE decided to start a new assessment system (Sistema de Información Sobre la Calidad de la Educación, SIMCE) managed by the Universidad Católica. The system was put in place to measure educational outcomes (mathematics, Spanish, social sciences and natural sciences) as well as affective development, and internal efficiency indicators. In 1988, new improved achievement tests were applied to 233,000 fourth graders in 5,600 schools; and a year later, in 1989, the tests were applied to 193,000 eighth graders in 4,600 schools. SIMCE articulated the following specific and very ambitious goals:
• assist the MOE in its normative work and in its supervision of the system;

• assist the regional and provincial authorities in supervision and technical support; and

• estimate quality in each educational institution, compare these results, look for explanatory factors, and evaluate the results of educational programs.

The main conclusions of SIMCE were:

• public municipal schools scored lower than “aided” schools and much lower than private schools;

• rural areas scored much lower than urban and suburban areas;

• there was a high correlation between socio-economic status (s.e.s.) and scores on the assessment; and

• an analysis of the 100 worst school confirmed that the common denominator was poverty and rurality.

The results were interpreted to mean that increased autonomy at the local level, as a means of introducing a private-sector "mentality" to all primary schools, might lead to increased achievement. The Government has been implementing a policy with this in mind.

A dissemination plan was prepared and the results were widely distributed. Each school received detailed results and parents were also informed of the school results. Booklets, audio-visual materials, and opinion surveys were used. The cost of the program per year was estimated at US$5 per student, or about 0.1% of the MOE’s budget.

Future plans include improving the cognitive instruments, adding more attitudinal instruments, analyzing the worst schools in greater detail, and undertaking more research.

Chile—Technical Aspects of the Assessment System

What was Measured. Chile opted for achievement rather than aptitude testing, and for criterion-referenced rather than norm referenced tests. However, in addition to cognitive items, Chile decided to measure individual development, including self-concept and self-esteem in relation to school experience, and also to measure opinions of students, parents and teachers on the educational process. It is not clear what Chile gained in terms of operational guidelines from these additional questions, except possibly greater political acceptance by those taking the examinations.

Chile opted mainly for a multiple choice format, but also included open-ended writing tests. Chile corrected only a random sample of 10% of the students’ writing samples—a potentially cost-effective approach to the problem of cost of open-ended questions. Information was not made available on any problems of reliability related to marking the open-ended questions, nor was any information available on how markers were trained and supervised.

Designing the Instruments. Chile has a centralized curriculum, applied in all regions of the country. For each grade level (4th and 8th) two criteria were used to identify the curriculum objectives that should be measured by SIMCE: that most schools covered the objectives and that they could be measured in an objective test format. Based on the official program for each subject matter, education specialists selected a list of fundamental objectives that were taught in most schools. Objectives that needed oral answers or demonstrations (i.e. the use of tools or machines) were excluded.

A committee formed by teachers and researchers identified items for each objective. Each item was analyzed in terms of its correspondence with the objective and the adequacy of style and distractors. Through
consensus items were rejected when these two aspects were not fulfilled. Finally, all selected items were tested in a pilot study among 3,500 students that had finished 4th and 8th grade the previous year; each item was answered by 350 students. Each item was evaluated by its difficulty level (items were accepted for the final set if they were correctly answered by at least 30% of the students), the internal reliability of the whole test (the item-test correlation had to be greater than 0.30), and the percentage of unanswered items (items that were not answered by at least 10% of the students were dropped of the final selection). The reliability of the final test, measured by the Kuder-Richardson approach, is higher than 0.90.

Background questions measured characteristics of the students and their families, teachers, classroom, directors, schools and the community where they are located. As noted above there may have been an excessive number of background questions.

With regard to booklet preparation and security, two parallel formats were assembled for each test. Packages were organized by destination identifying region, city and school, and a list was provided to control the number of tests sent to each destination and the number of tests returned. Overall security seems not to have been a problem.

Administering the Assessment. In Chile all students in fourth and eighth grades took the mathematics and Spanish tests, with the exception of isolated schools which were excluded from the assessment. Total coverage was 90%. The social science and natural science tests were applied to a sample of students using a stratified sampling frame that controlled for socioeconomic level; geographical location, including rural urban location; and type of school (public municipal, subsidized private, and non-subsidized private). This scheme will be kept for future applications. It was decided to assess all students in order to provide feedback to all teachers in the key areas of mathematics and Spanish. This approach led to relatively high costs and could have been replaced, at least partially, by sample surveys. For the natural sciences and social sciences, a random sample of 10% of the students was selected. In addition, although the writing test was applied to all students only a random sample was corrected.

The field work lasted a week but all the tests are administered in two days, Wednesday and Thursday. All 40 Direcciones Provinciales participated in the process with the collaboration of the supervisors (748) who selected and train the test monitors and interviewers (9,000, selected among school teachers). Optical scanners at the Universidad Católica were used to score the multiple choice tests. The results of the writing scores were mechanically added to the other test scores.

At the Universidad Católica the information was validated and print outs produced at the student, school, municipal and regional level. Using LOTUS and SPSS, basic frequencies are obtained, as well as cross-tabulation at different levels including students' socioeconomic status, school modality and school administration, and region. Little or no multivariate analysis was carried out.

Disseminating the Results. In Chile reports of the average achievement (on a class by class basis) for each test were sent to central and regional educational authorities, supervisors, school directors, teachers, municipal authorities and parents. The reports summarized achievement by each learning objective, identifying the percent of students that reached each objective. To interpret the results two mechanisms are used: an explanation of how to compare schools was provided in the reports and meetings using audio-visual materials were held with supervisors, directors and teachers.

A manual on pedagogical implications of the tests was provided by SIMCE. Only 35% of teachers reported using the manuals. On a scale
from 0 to 10 points from not useful to highly relevant, teachers scored the results at an average of 5.7, supervisors at 7.6 and provincial authorities at 7.7.

Research and cost effectiveness analysis was not undertaken by SIMCE in spite of the potentially rich results. However the fact that private institutions did far better than public institutions led to a major debate on re-organizing the education system.

Managing the Assessment. In Chile SIMCE has been under the direction of the Universidad Católica, and its director and the academic and administrative directors are university faculty members. An interdisciplinary group of educators, psychologists and system engineers from the Universidad Católica, MOE and consultants from UNESCO implemented SIMCE. The university was responsible for test design, data collection and systematization, data analysis, and data reporting. The Government is planning to transfer the system to the MOE (with technical assistance from the Universidad) and a proposed Bank loan would include financing to help institutionalize the system. This approach runs a risk of excessively bureaucratizing the system and will reduce flexibility. It might be better to continue to devolve responsibility for daily operations to an agency outside MOE like Universidad Católica. However, dissemination will have to remain the overall responsibility of MOE. In any event there is a need for a training program to improve the skills of the people involved in SIMCE as well as to promote the interdisciplinary work of psychometricians, system engineers and educators.

Costa Rica—History and Background

Two factors gave origin to the assessment of primary education in Costa Rica. First, since the educational reform in the sixties, teachers and parents thought that the quality of primary education had been deteriorating as a result of the automatic promotion policy and the elimination of the entrance exam to secondary education. Second, results from research conducted between 1981 and 1986 by the Instituto de Investigaciones para el Mejoramiento de la Educación en Costa Rica (IIMEC), of the Universidad de Costa Rica, showed low education achievement among primary students. In 1986 MOE asked IIMEC to test students from 3rd, 6th, 9th grades in basic education and the last year in secondary education in mathematics and Spanish. The test for the last year of secondary education was designed for purposes of certification and was also used for university entrance. In 1986 and 1987 the universe of students were tested. In 1989 a stratified sample was tested.

The objective of these surveys was never clearly articulated. However it was clear that the MOE wished to make the public aware of the low level of education and to have a national debate on the question. A second, but not articulated objective, was to prepare the public for the re-introduction of summative examinations at the end of the sixth and ninth grades.

The results of the assessment were as follows:

* the achievement scores were low in comparison with the nation-wide curriculum objectives. However, the low scores can be explained in part by the fact that students had little interest in scoring well because tests were known to be diagnostic and not summative and because teachers themselves were not adequately urged to motivate the students;

* there was widespread lack of basic knowledge in the fundamental areas of Spanish, math, science, and social studies;

* private institutions did best, followed by private institutions with some public support. Public institutions did far worse than the private and semi-private schools;
urban schools did much better than rural schools, and larger schools generally did better than smaller schools;

day time academic secondary schools did best, followed by technical schools, agricultural schools, and night-time academic schools; and

achievement in later grades seemed to be lower than achievement in the earlier grades in comparison with the official curriculum.

Little thought was given by MOE to feedback to individual teachers and a long term assessment plan was not articulated. IIMEC undertook on its own to provide to individual schools, but MOE staff did not use the information, possibly because the initiative for the assessment was very strongly associated with the Minister. IIMEC sent information to each school as well as to the press on its scores. In addition IIMEC sent information on the extent to which specific curriculum objectives were being met through the test. The cost per student varied from US$1.34 to US$7.69. The total cost was generally around 0.1% of the MOE’s budget.

Subsequently MOE decided to abandon the idea of a centrally prepared and administered test in favor of tests developed by district on the basis of national criteria. The objective of these tests is to establish a minimum level of knowledge necessary for graduation from primary and secondary school, and to improve quality through giving more responsibility to the student, parent, and teacher. These district level tests in the last year of primary education count 50% in the final grade a student receives, and therefore the implicit objective of the initial assessments appears to have been met. They are prepared at the district level by groups of teachers on the basis of national guidelines as well as locally defined "needs." However the district level tests are not scored by machine and therefore systematic feedback is not possible, nor is there an objective way of comparing district by district results. There is also a potentially serious problem of security since many teachers participate in the preparation of the test. For the future MOE has proposed strengthening its own capacity for testing and measurement and using IIMEC for technical assistance rather than as a program implementor. In April 1991, MOE began planning a new assessment system, in connection with a proposed World Bank project.

Costa Rica—Technical Aspects of the Assessment System

What Was Measured. Costa Rica’s assessment was criterion-based, measuring what students learned with respect to the intended curriculum. Initially only achievement in mathematics and Spanish was covered. In the second year, natural sciences and social studies, as well as French and English for 9th grade, were added. Only a small component of the Spanish test included an essay, with all the other tests based on a multiple choice format.

Designing the Instruments. Costa Rica has a centralized curriculum. The objectives of the test were selected on the basis of the official study programs. A group of 500 teachers ordered the curriculum objectives from the most to the least important. Specific cognitive items corresponding with test objectives were selected by teachers and MOE personnel; a final group of 10 judges verified if they had content and instructional validity before a final selection was made. Three measures of reliability, including the Kuder-Richardson score, were used to establish the quality of the items. Standard background questions identifying school and family characteristics were included.

Editing of tests was troublesome because MOE lacked the technology to revise texts. Test printing and test storage was conducted by private companies and security problems were observed, possibly lowering the overall validity of the tests. With regard to pre-testing, parallel tests, each with three items by objective, were
applied among a sample of students to identify item discrimination and difficulty.

**Administering the Assessment.** In 1986, the Costa Rican test covered the entire population of students. In 1987 the test covered all students, with the exception of multi-grade single-teacher schools (around 9%). In 1989 a stratified sample of students was used for the application of the tests.

IIMEC and MOE personnel together with hired teachers gave the tests. Teachers were not permitted to give the tests in their own schools. The primary level assessments were scored by optical scanner. The tests at the end of secondary level were hand scored. IIMEC undertook item analysis and descriptive analysis, including simple cross-tabulations by student and school characteristics.

**Disseminating the Results.** In Costa Rica all schools got a summary of their students’ average achievement and a comparison with other schools. Near 70% of the teachers reported that they had seen the results, but only a small percentage of supervisors and teachers reported that they had used the results to improve the way they teach in the classroom. Reports were produced for the school, regional and central authorities. Based on the test results, MOE argued in a variety of public forums for increased investment in primary education and for increased parental involvement. MOE used the results of the initial tests to develop an in-service teacher training program which has not yet been implemented. No significant follow-up research has been undertaken, with the exception of some small scale analysis by IIMEC. With regard to feedback, IIMEC benefitted from experience of the first three years of the program. MOE will utilize IIMEC in the design of the next stage of assessment to be financed under a proposed World Bank loan.

**Managing the Assessment.** In Costa Rica MOE contracted IIMEC to develop the achievement tests and the statistical analysis, as well as the procedures for packaging, distributing and test application. MOE personnel conducted the actual field work. The Department of Evaluation at MOE does not have the technical capacity nor the resources to conduct assessments. MOE now proposes to develop the "Sistema de Información Sobre la Calidad de la Educación Costarricense", SICEC, based on the IIMEC experience. MOE does not have trained technicians to conduct an assessment and to establish a research operations including pedagogical analysis of the tests. MOE is expected to focus on strengthening IIMEC and establishing within MOE only a very small high quality assessment unit (2-5 staff).

**Mexico—History and Background**

Periodic assessments of the quality of the education in Mexico have been conducted for the last 20 years, including experiences with the International Association for the Evaluation of Educational Achievement (IEA), and the ECIEL study of educational achievement. Since 1972 Mexico also has had entrance exams to secondary education. At that time 361,000 students took the secondary entrance exam. The number has since increased to more than 1,200,000 in the 1980s.

The Dirección General de Evaluación y de Incorporación y Revalidación (DGEIR) has undertaken several studies (varying in sample coverage, academic subject evaluated, periodicity, depth of analysis, etc) designed to assess the quality of different education levels. Some of the DGEIR studies assessing the quality of education piggybacked on information from the entrance exams. Although a methodology has been developed as a result of the experience, due to the economic crisis, little was done during the 1980s because DGEIR lost much of its specialized human resources due to salary cuts. Today most of the staff lack training in relevant areas of assessment and evaluation. Currently, with the support of the World Bank to recuperate the capacity that deteriorated during the 1980s, a technical assistance program
is being design with the Educational Testing Service (ETS) to strengthen DGEIR capacity to conduct assessment.

As a result of the Education Modernization Program, 1989-1994, DGEIR has been seeking to systematize its assessment in all 31 states, among others through the study " Evaluación del Aprendizaje en Educación Prescolar, Primaria y Secundaria (EAEPPS)" , which is part of the Sistema Nacional de Evaluación Educacional (SNEE). SNEE objectives seek to evaluate the learning of subject matters at different grade levels, highlight regional differences across states, and evaluate the effectiveness of pedagogy across schools. The results will identify deficiencies and areas of excellence, which will be used by the Secretaria de Educación Pública (SEP) to reallocate resources and refine its curriculum, teaching and educational infrastructure improvement program. As dissemination of the results of prior experiences was quite limited, SNEE has started an awareness-raising program among parents, students, but mainly teachers and local authorities, so that they understand the assessment process.

Similarly to other Latin American countries, the results of the Mexican experience show that the majority of the students have low achievement (i.e in a national study conducted in 1988 among primary students, 69% scored a grade of 6 or lower on an academic performance scale from 0 to 10), and that achievement is significantly related to student socioeconomic status and school urban-rural location. At DGEIR the basic level of analysis is distribution of frequencies so little information is available on the factors associated with low achievement. It should be mentioned, however, that DGEIR conducts pedagogical analysis of the results and provides feedback to the schools on detected problems.

The cost per student, around US$35, is higher than the ones in other Latin American countries because in the Mexican case the survey is also administered to parents, teachers and headmasters. The current budget of DGEIR is less than 0.1% of the SEP budget.

Mexico—Technical Aspects of the Assessment System

What Was Measured. In Mexico a typical assessment study for primary education measures cognitive achievement and ability among a sample of sixth graders (the last grade of primary education). For example the CEIS-80 study, a complement study to entrance exams in 1980, measured cognitive achievement in Spanish, mathematics, social and natural sciences. The EAEPPS measures cognitive and affective aspects of students completing preschool, cognitive achievement in Spanish, mathematics, social and natural sciences for students in each grade of primary school, and those four areas plus English for students in each grade of secondary education. Nearly all tests measure cognitive achievement (a few include aptitude measures), all are criterion-referenced, and all use multiple choice formats.

Designing the Instruments. For the EAEPPS in Mexico, a group of experts reviewed the study programs used during the last 15 years to identify the contents present during this period. According to the criteria of DGEIR experts some relevant contents were included even if they were not present in all study programs during the period. The result was a desired academic profile for each grade from preschool to secondary. Once the test objectives were identified, items were selected by expert consensus. Item analysis, including degree of difficulty, reliability and validity has been undertaken but not as a general rule, with a result that, according to some experts in the DGEIR, the results of some studies are not valid. Recently the situation has deteriorated as DGEIR has lost funding as well as expertise for item design and analysis. Most studies include standard sets of background variables, including characteristics of the school, teacher, student and family background, which have been tested and
confirmed over time. DGEIR has standard procedures to administer the test and to implement the field work required to collect and systematize the information. Pilot testing is regularly undertaken.

Administering the Assessment. In Mexico, with the exception of a test of graduates of teacher training schools, all assessment studies are based on samples which vary in terms of size, geographical area, and sampling design. DGEIR officials report that, compared to five years ago, today they have inadequate economic and human resources to select valid samples. Generally DGEIR uses large samples covering at least a couple of states and Mexico city. The typical unit of selection is the school and a random sample is used to select students in the classrooms. The tests are typically given by trained interviewers that are hired for that particular purpose. The computational capacity available to score and systematize the data is limited. There is access to software for statistical analysis but it is inadequately used. Test items are validated. The typical statistical analysis is that of frequency distributions, but correlational analysis has been used at times.

Disseminating the Results. In Mexico, reports on studies are distributed among education authorities, supervisors and teachers. During the last two years DGEIR technicians have organized meetings to discuss the results. However there is very little educational analysis and few, if any, pedagogical recommendations, are made. Local researchers have sometimes used the results of studies to undertake research. However the DGEIR does not undertake such research.

Managing the Assessment. In Mexico, despite its far-reaching mandate, DGEIR's capacity to conduct assessments has been decreasing and it is losing qualified technicians. Salaries have deteriorated and the private sector has attracted many staff members. Currently DGEIR has no sampling experts, and statistical analysts are badly out of date on techniques for item design and statistical analysis. Under a proposed World Bank loan, DGEIR's capacity will be strengthened.

Colombia—History and Background

As noted above, Colombia is of interest not so much because it has undertaken assessments but because it has a well structured and organized system of university entrance examinations which could be built upon for purposes of assessment. The system in Colombia began in 1959, through a vocational guidance program under the auspices of the National University Association to help secondary education graduates select university studies. The program included testing in topics such as verbal ability, mathematical reasoning, personality traits, general knowledge and university study preferences. In 1962, the Educational Testing Service (ETS) at Princeton trained faculty members from several Colombian universities on test design and analysis in order to improve the quality of the tests that at the time were being implemented among high school graduates who wanted to follow higher studies. About ten universities participated in the program. In 1968, the Colombian Servicio Nacional de Pruebas (SNP—Colombian National Testing Service) was established to administer a university entrance exam for all high school graduates seeking university entrance. In 1984 SNP initiated a public secondary education entrance exam in Bogotá in order to assign limited places. Up to 1986 these secondary entrance exams measured aptitudes (verbal, mathematical and abstract reasoning). From 1987 cognitive tests were included in math, Spanish and natural and social sciences. All tests are multiple choice with four alternative answers. Annex 4 of this report provides samples of the background questionnaire and instructions to students of one of these highly professionally prepared and administered examinations.

Like the other countries of the Region, in Colombia the research capacity to analyze the
results of the tests is very limited. Typically only frequencies and cross-tabulations are presented, and occasionally a more detailed analysis is carried out by contracted research groups outside of the SNP. The main results reported by recent studies could be summarized as follows:

- students from private schools get better scores, mainly in math and social studies;
- male students do better than female students;
- day time schools do better than night schools;
- technical schools get the higher scores followed by academic schools, agricultural and commercial schools get the lowest scores;
- significant differences are observed by geographical regions and by urbanization level; and
- younger students and students whose parents have higher education get better scores.

In 1990 the cost of the university entrance exams was near US$4 per student and each of them was charged with US$2.66 recovering a significant proportion of the total cost. The exam for secondary entrance exam was US$2 per student, 10% of which was charged to the student. In any event, the total cost of both tests (US$1.4 million) in 1990 was less than 0.01% of the Ministry of Education budget.

Colombia--Technical Aspects of the Assessment System

What Was Measured. In Colombia the CNTS secondary school leavers examinations, designed to select students for higher education, measure verbal and mathematical aptitude as well as achievement in biology, chemistry, physics, Spanish, mathematics and social studies (including philosophy, geography and history).

All students take these eight tests plus an elective from the following list: abstract reasoning, spacial relationships, mechanic analysis, English, pedagogical knowledge, agriculture, accounting, French, metalwork, and electricity. The CNTS tests are similar for primary school leavers in the Bogotá area (selecting for entrance to public secondary schools). Because their purpose is to select students, CNTS tests are norm-referenced; the emphasis is on the relative position of the students rather than on ability or on mastery of the curriculum content. However the content of the achievement tests is based on the primary or secondary curriculum. All tests are multiple choice with five alternative answers in the case of the university entrance exam and four alternative answers for the public secondary school entrance exam.

Designing the Instruments. Because Colombia has a centralized curriculum, applied in all regions of the country, the university entrance exam is national and is based on secondary school curriculum objectives. The methodology to identify the test objectives has been the same since 1968: SNP coordinates a group of experts, including teachers and curriculum programmers, to identify curriculum objectives according to Bloom's taxonomy. For each subject, once a list of fundamental objectives is established, a group of teachers and measurement experts identify items to measure memory, understanding, application, analysis and evaluation for each objective. Each year some of the proposed new items are included in the exams to test their reliability and validity. Measurements of degree of difficulty and degree of item-test co-variance are used to determine whether to include a new item in the item bank. The tests ask for identification of the school, including geographical location, type of administration, nature, and modality, and socio-economic characteristics of the student, including gender, sex, age, marital status, family income, father and mother education and occupation. The SNP has developed a series of manuals on procedures to organize and implement the
exams. There are manuals for the field coordinator or delegate, classroom coordinators, and instructions for room supervisors.

**Administering the Assessment.** In Colombia the university exam is mandatory for all secondary graduates seeking university studies. The secondary level tests are required for students wishing to attend a public secondary school in Bogotá. For these reasons the tests cover a universe rather than a sample. The number of students taking the university entrance exam has increased from 28,000 in 1968 to 150,000 in 1980 and 301,000 in 1990. The number of students taking the public secondary entrance exam in 1984 was 60,000 and in 1990 74,000.

The SNP has developed a comprehensive security system, starting with the management of the item bank, and including printing, packaging, transportation to school, distribution in the classroom, gathering of answered tests, and transportation back to the SNP headquarters. A private security firm, Thomas de la Roux, is responsible for most of these tasks.

Optical scanners are used to process the tests. Once the tests are validated, the results by student and schools are obtained by software developed by the SNP. SNP undertakes simple statistical analysis, including frequency distribution and bi-variate analysis to contrast tests results by some school and student socioeconomic characteristics. More efficient software like SPSS or SAS is not yet used and could help to improve the extent and quality of the current statistical analysis.

**Disseminating the Results.** In Colombia the SNP is highly efficient in producing and disseminating test results. Within six weeks of the test, SNP reports results to the student and her/his respective school. The SNP informs the general public about the results at the school level and annually the national press publishes a list of the best 100 schools. Because of inadequate statistical analysis the reports produced by the SNP are very descriptive and of little use for educational policy. The original objectives of vocational guidance has been forgotten and the system is only used for selection purposes. There is no feedback to schools on pedagogical aspects, since the system is highly focussed on the selection objective. SNP evaluates each test application. A special format called "Closing of Program" has been developed for this purpose. It includes statistics on numbers of schools, students, etc., costs, dates, description of problems during the application, and proposed solutions.

**Managing the Assessment.** In Colombia, the support that MOE gave to the CNTS when it started in terms of resources and organization, and the effort in shaping a team of skillful technicians has been the key to the successful experience of the CNTS. Out of the 63 professional staff working in CNTS, 28 are responsible for item design and analysis and research. While item design and analysis is well done, almost no research is conducted. There is no education analysis of the results, no feedback to the educational authorities, and therefore no explicit effect on educational policy. It would seem that CNTS would be an excellent candidate for the feedback experiment undertaken in Kenya and described in Somerset's report on "Examination Reform: the Kenya Experience."

**The LAC Experience: Summary and Recommendations**

*Chile* has had the most successful experience with assessments. Chile initiated its program in 1978. After initial problems, including resistance from teachers and students, the program, in place since 1988, has successfully measured learning for universes among fourth and eighth graders. The aim of the program has been to affect educational policy and to strengthen teaching practice. The program included a strong dissemination effort involving civil servants, teachers, and parents. The program has influenced government policy and there is some evidence that it has affected
classroom practice. The relative success of the Chile program appears to be a result of (a) the high quality of its staff, and (b) a strong focus on providing feedback directly to teachers, schools and districts and on informing the general public of the nature and role of assessment program. However, the program has not included research on causes of school failure and has relied almost exclusively on high cost censuses rather than on samples. Furthermore, the relationship between the Ministry of Education (MOE) and the Catholic University implementing the program up to now requires clarification.

Costa Rica initiated its program in 1986 and 1987 and assessed the universe of third, sixth and ninth graders. The program did not have a clear articulated goal, but its implicit goals included (a) using the assessments as a tool to argue in the public arena for additional funding for primary education, and (b) convincing the public of the need to re-introduce the use of partly standardized tests for certification of secondary school graduates. The assessment did not include the objective of using the assessment as a direct tool to improve classroom practice. Feedback to schools and teachers has been ad-hoc. Costa Rica has a limited human resource base, especially in Government, for this kind of activity.

The Government of Mexico has undertaken numerous assessments and evaluations over a twenty year period. It has also analyzed the national secondary school entrance examination for assessment purposes. The agency responsible for most of this work recently lost many of its good staff and has had an inadequate budget. It has been disseminating results, but on an excessively theoretical and general basis. It is now hoping to strengthen its staffing and is planning a stronger awareness raising program for parents, teachers and school authorities.

Colombia has an excellent autonomous agency undertaking university entrance examinations and examinations for entrance to public secondary schools in the Bogotá metropolitan area. Colombia’s exams are a model of modern computerized test preparation and scoring. The challenge for Colombia is to build on this capacity through utilizing selection examinations for purposes of assessment, similar to that of Kenya, as well as to develop a primary school level assessment system.

None of the four countries has adequately incorporated research into assessment and testing. This means that much of the value of their effort is being lost because of a lack of additional complementary inputs. Furthermore there is an often unnecessary tendency to test the universe of students rather than to use less expensive samples. Using sample survey methodology will depend on good statistical expertise which is a scarce commodity. Furthermore these four countries have not put enough effort into dissemination of test results. Chile has the best record but could still do much more to have an impact on classroom behavior. Mexico’s dissemination efforts, while widespread, have been excessively general and theoretical and data on a per school or per district basis have not been systematically provided. Costa Rica’s dissemination efforts were undertaken on an ad-hoc basis by a testing agency outside Government. Annex 5 provides samples of dissemination in the form of statistical tables from Costa Rica and Chile. These should be considered as first stages in developing dissemination systems. The Kenya example (Annex 1) is a far more sophisticated example which should be a goal for most dissemination programs. Annex 6 provides a good example from Colombia on information provided on testing to teachers.

Coupling these four cases with the technical and managerial issues described in Chapter III, Latin American countries seeking to establish or strengthen educational assessment systems will need to pay particularly close attention to the following elements:
● From the start Latin American educational authorities will need to define clearly the objectives of the assessment, including specifically how the assessment is expected to lead to improved classroom practice. They will also need to prepare a five-year implementation and cost plan, and act to ensure continuous and full support from the highest government authorities.

● From the start the managers of the assessment will need to design a full dissemination plan. The managers will need to identify the various clienteles who will use the assessment, hire expert writers and reporters for dissemination, and plan for in-service training programs of teachers based on the assessments. A strong effort will need to be made to educate all concerned parties as to the fact that an assessment is not designed to award or punish, only to assist, and that low scores in a region, city, or school, may not reflect on the teachers since there are many operating external factors.

● Latin American countries should seriously consider keeping the testing department inside government very small. In principle most of the assessment design and implementation work should be contracted out to an independent, stable non-profit agency. If such an agency does not exist, then efforts should be made to build one up over time.

● Latin American countries will need to seriously consider using sample surveys rather than censuses, especially when the objective is to measure performance of groups of students, schools, and the system as whole. But samples had best not be undertaken if the sampling methodology is inadequate.

● Assessment programs will need to include funds for research, as well as agreed upon cooperative programs with local and/or foreign independent non-profit research institutions.

● Latin American governments will need to finance the training of psychometricians and statisticians. Contracting with non-profit agencies will help to ensure that this expertise remains available to government, since it is difficult for government to retain qualified staff with these scarce skills.

● Countries such as Colombia with well managed selection examinations will need to consider utilizing these examinations for assessment purposes.
REFERENCES


Sample Feedback from Kenya

In analysing the science questions in the 1979 General paper, an interesting fact was discovered about differences in performance between boys and girls. As you may remember from p 26 of this Newsletter, the performance gap between boys and girls was higher in science than in any other subject. In English, the average gap was only 3.6%; in mathematics, it was 7.1%; but in science it was as high as 7.7%.

When the science questions were analysed separately, however, it was found that girls performed quite well in most of the reasoning questions. For example, the question about bases and acids quoted above (no. 61) was answered correctly by 46.2% of boys and 43.6% of girls, so the performance gap was only 2.6%.

By contrast, girls tended to do badly in most of the questions which gave an advantage to pupils who had carried out practical observations and experiments. For example, the question about how a card would hang from a nail (no. 84) was answered correctly by 58.6% of boys, but only 40.5% of girls. Thus the performance gap was as high as 18.1%.

When the scores of boys and girls in all reasoning and all observation questions were compared, the results were as follows:

<table>
<thead>
<tr>
<th>Questions testing</th>
<th>Average mark</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Boys</td>
<td>Girls</td>
</tr>
<tr>
<td>(a) mainly reasoning</td>
<td>46.4%</td>
<td>42.0%</td>
</tr>
<tr>
<td>(b) mainly observation or experiment</td>
<td>56.5%</td>
<td>42.4%</td>
</tr>
</tbody>
</table>

As you can see, the average performance gap in questions which tested mainly reasoning was only 4.4% which is only a little higher than the gap for English. By contrast, in the questions which tested mainly the results of observations or experiments the gap was as high as 14.1%.

These results suggest strongly that the main reason why boys perform better than girls in science is not that they have superior reasoning ability, but rather that they have had more experience in carrying out observations and experiments. Make sure that when you are doing practical work in science, the girls take just as active a part in the observations and experiments as the boys. If you find that girls tend to be passive when

One of the questions in the science section of the paper which tested a terminally-relevant skill deserves special mention because it was answered so poorly:

62. Said goes to the market to buy a fish. He notices that the pointer of the balance is set as shown in Diagram P. When the fish is placed in the pan, the pointer is as shown in Diagram Q. Which one of the following is the most likely weight of the fish?

A. 1 kg. 100 g.
B. 1 kg. 200 g.
C. 1 kg. 300 g.
D. 1 kg. 400 g.

Only 18.5% of rural Candidates, and 20.1% of Nairobi candidates (excluding those in high-cost and private schools) correctly worked out that the most likely weight of the fish was 1 kg. 300 g. More than 40% of candidates in both samples chose 1 kg. 400 g. This, of course, is the candidates ignored the information given in the first diagram (P), which shows that before the fish was placed in the pan the pointer was set to 100 grams. Thus the fish probably weighs 100 grams less than the reading shown in Q. If Said made the same mistake as these candidates, he probably paid too much for his fish!

When a customer buys sugar, meat or fish from a shop, or when a farmer sells his pyrethrum, coffee or maize, he needs to be able to check that the weighing is done accurately. During geography and science fields trips, pupils should become familiar with balances and scales of as many different types as possible. Pupils can observe weighing being carried out in places such as coffee factories, post offices and tea 'centres' (places where tea farmers sell their green tea leaves) as well as shops and markets.
**Table 5-8. Vocational Achievement Scores as a Function of School Type and Out-of-School Characteristics, Colombia, 1981 Cohort**

<table>
<thead>
<tr>
<th>Independent variable</th>
<th>Agricultural achievement</th>
<th>Commercial achievement</th>
<th>Industrial achievement</th>
<th>Social services achievement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coefficient</td>
<td>t-value</td>
<td>Coefficient</td>
<td>t-value</td>
</tr>
<tr>
<td><strong>Out-of-school factors</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>0.950*</td>
<td>3.66</td>
<td>-0.804*</td>
<td>3.20</td>
</tr>
<tr>
<td>Age</td>
<td>0.141</td>
<td>1.81</td>
<td>0.153*</td>
<td>2.10</td>
</tr>
<tr>
<td>Family owns business</td>
<td>-0.497</td>
<td>1.34</td>
<td>-0.497</td>
<td>1.43</td>
</tr>
<tr>
<td>Household books</td>
<td>0.007*</td>
<td>2.59</td>
<td>0.003</td>
<td>1.38</td>
</tr>
<tr>
<td>Father's education</td>
<td>0.006</td>
<td>0.15</td>
<td>0.089*</td>
<td>2.50</td>
</tr>
<tr>
<td>Log of family income</td>
<td>0.120</td>
<td>1.40</td>
<td>0.255*</td>
<td>3.18</td>
</tr>
<tr>
<td>Father is employer</td>
<td>-0.593</td>
<td>1.30</td>
<td>0.346</td>
<td>0.81</td>
</tr>
<tr>
<td>Father blue-collar</td>
<td>-1.427*</td>
<td>3.55</td>
<td>0.012</td>
<td>0.03</td>
</tr>
<tr>
<td>Father nonmanual employee</td>
<td>-0.803*</td>
<td>2.73</td>
<td>0.041</td>
<td>0.15</td>
</tr>
<tr>
<td>Verbal aptitude</td>
<td>0.233*</td>
<td>15.20</td>
<td>0.132*</td>
<td>9.17</td>
</tr>
<tr>
<td>Math aptitude</td>
<td>0.001</td>
<td>0.05</td>
<td>0.051*</td>
<td>3.74</td>
</tr>
<tr>
<td><strong>Curriculum</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control specialty</td>
<td>11.500*</td>
<td>19.53</td>
<td>8.851*</td>
<td>23.13</td>
</tr>
<tr>
<td>Constant</td>
<td>33.235*</td>
<td>15.95</td>
<td>32.249*</td>
<td>16.42</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.13</td>
<td>0.25</td>
<td>0.09</td>
<td>0.15</td>
</tr>
<tr>
<td>Sample size</td>
<td>5,838</td>
<td>5,838</td>
<td>5,838</td>
<td>5,838</td>
</tr>
</tbody>
</table>

**Note:** Each dependent variable is the score on a specialization test. The INEM specialty dummy refers to students enrolled in the INEM track corresponding to the particular test in that column. The control specialty variable represents students in the non-INEM specialty track. Regression coefficients are to be interpreted with reference to the omitted category—that is, all students not following that particular subject.

a. Statistically significant at the 5 percent level or better.

### IMPACT OF PRESCHOOL, SCHOOL AND TEACHER VARIABLES ON READING ACHIEVEMENT IN PERU

6th Grade Students  
(N=324)

<table>
<thead>
<tr>
<th>S.E.S. AND BACKGROUND:</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>student sex</td>
<td>-0.34</td>
<td>-</td>
<td>-</td>
<td>19.8%</td>
<td>17.0</td>
<td>***</td>
</tr>
<tr>
<td>student age</td>
<td>-1.10</td>
<td>2.24</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mother’s education</td>
<td>-0.20</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>father’s occupation</td>
<td>0.31</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>newspapers at home</td>
<td>0.43</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCHOOLS LOCATION</td>
<td>-0.62</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**SCHOOL VARIABLES**

1. School Management  
   - percentage of time correcting exercises 0.26 - -  
   - monthly rate of absenteeism -1.85 2.51 ***  
   - criterion for class grouping -3.04 2.89 ***  
   - remedial programs 1.87 1.95 **  
   - percentage of time explaining -2.16 3.11 ***

2. Teacher  
   - teaching hours per week 0.89 2.83 *** 32.9% 14.2 ***

3. Physical Facilities  
   - building age 0.81 3.94 ***  
   - electric light 4.89 3.66 ***  
   - library 7.17 2.93 **  
   - presence and use of individual aids 1.64 1.84 *  
   - percentage of students with writing facilities 0.93 2.15 **

<table>
<thead>
<tr>
<th>ALL VARIABLES</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>43.0%</td>
<td>15.36</td>
<td>***</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**t ="t" coefficient.**  
**s.l = significance level.**  
***** = statistically significant at 0.01.**  
** ** = statistically significant at 0.05.**  
** * = statistically significant at 0.10.**

**Source:**  
A. M. Arriagada Determinants of Sixth Grade Student Achievement in Peru. Washington, D. C. The World Bank, Education Department, 1983.
Table 5: School type effects on achievement and social class differentiation

<table>
<thead>
<tr>
<th>Independent variable</th>
<th>Mathematics</th>
<th></th>
<th>Portuguese</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeff.</td>
<td>t-stat</td>
<td>Coeff.</td>
<td>t-stat</td>
</tr>
<tr>
<td><strong>Fixed</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>-1.48***</td>
<td>-7.06</td>
<td>0.58**</td>
<td>3.02</td>
</tr>
<tr>
<td>Age</td>
<td>-0.37***</td>
<td>-5.84</td>
<td>-0.48**</td>
<td>-8.24</td>
</tr>
<tr>
<td>Family size</td>
<td>0.05</td>
<td>0.27</td>
<td>0.05</td>
<td>1.11</td>
</tr>
<tr>
<td>Working hours</td>
<td>-0.03***</td>
<td>-4.65</td>
<td>-0.02**</td>
<td>-3.62</td>
</tr>
<tr>
<td><strong>Mean achievement</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>13.92***</td>
<td>22.29</td>
<td>16.81***</td>
<td>43.71</td>
</tr>
<tr>
<td>Average SES</td>
<td>4.71***</td>
<td>4.74</td>
<td>3.82***</td>
<td>6.42</td>
</tr>
<tr>
<td>Federal technical</td>
<td>9.28***</td>
<td>5.20</td>
<td>3.68**</td>
<td>3.44</td>
</tr>
<tr>
<td>SENAI</td>
<td>-1.52</td>
<td>-0.63</td>
<td>-0.84</td>
<td>-0.58</td>
</tr>
<tr>
<td>Teacher training</td>
<td>-1.37</td>
<td>-1.06</td>
<td>-0.09</td>
<td>-0.12</td>
</tr>
<tr>
<td>Private</td>
<td>3.20*</td>
<td>2.12</td>
<td>-0.20</td>
<td>-0.22</td>
</tr>
<tr>
<td><strong>SES achievement gap</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base</td>
<td>0.11</td>
<td>0.54</td>
<td>0.21</td>
<td>1.26</td>
</tr>
<tr>
<td>Federal technical</td>
<td>0.68</td>
<td>1.31</td>
<td>-0.25</td>
<td>-0.57</td>
</tr>
<tr>
<td>SENAI</td>
<td>-0.69</td>
<td>-1.07</td>
<td>-0.37</td>
<td>-0.70</td>
</tr>
<tr>
<td>Teacher training</td>
<td>-0.35</td>
<td>-0.80</td>
<td>-0.28</td>
<td>-0.74</td>
</tr>
<tr>
<td>Private</td>
<td>0.02</td>
<td>0.06</td>
<td>0.22</td>
<td>0.72</td>
</tr>
</tbody>
</table>

* p < .05, ** p < .01, *** p < .001

**Sample Background Questions for ECIEL Study**

**SCHOOL AND TEACHER VARIABLES QUALIFYING FOR THE FINAL REGRESSIONS IN EACH ECIEL COUNTRY**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Countries in Which Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Map in classroom</td>
<td>Bolivia, Brazil</td>
</tr>
<tr>
<td>Teacher cabinet</td>
<td>Colombia, Mexico</td>
</tr>
<tr>
<td>Number of library books on loan</td>
<td>Argentina, Mexico, Brazil</td>
</tr>
<tr>
<td>Times library used per week</td>
<td>Paraguay</td>
</tr>
<tr>
<td>Times library used per month</td>
<td>Colombia, Argentina, Mexico</td>
</tr>
<tr>
<td>How many students attend library per month</td>
<td>Peru</td>
</tr>
<tr>
<td>Hours lab used per week</td>
<td>Argentina</td>
</tr>
<tr>
<td>Hours per work week</td>
<td>Argentina, Bolivia, Mexico</td>
</tr>
<tr>
<td>Total payroll (US$) teaching staff</td>
<td>Mexico</td>
</tr>
<tr>
<td>Age of school buildings</td>
<td>Bolivia, Peru, Paraguay</td>
</tr>
<tr>
<td>Land area of school</td>
<td>Mexico</td>
</tr>
<tr>
<td>Estimated value of building</td>
<td>Argentina</td>
</tr>
<tr>
<td>Students pay into school fund</td>
<td>Argentina</td>
</tr>
<tr>
<td>School gives free materials</td>
<td>Paraguay</td>
</tr>
<tr>
<td>Annual cost of materials given to students</td>
<td>Paraguay</td>
</tr>
<tr>
<td>Total number of primary-school classes</td>
<td>Paraguay</td>
</tr>
<tr>
<td>Highest monthly absence rate</td>
<td>Argentina, Mexico</td>
</tr>
<tr>
<td>Absence rate last month</td>
<td>Peru</td>
</tr>
<tr>
<td>Expenditure per student</td>
<td>Argentina</td>
</tr>
<tr>
<td>Principal qualification level</td>
<td>Paraguay</td>
</tr>
<tr>
<td>Years as school principal</td>
<td>Peru</td>
</tr>
<tr>
<td>Years principal of this school</td>
<td>Colombia</td>
</tr>
<tr>
<td>School used</td>
<td>Bolivia</td>
</tr>
<tr>
<td>Hours per day school is open</td>
<td>Peru</td>
</tr>
<tr>
<td>Number of school shifts</td>
<td>Peru, Mexico</td>
</tr>
<tr>
<td>Number of hours per day</td>
<td>Colombia, Argentina, Brazil</td>
</tr>
<tr>
<td>School has telephone</td>
<td>Bolivia</td>
</tr>
<tr>
<td>School has auditorium</td>
<td>Mexico</td>
</tr>
<tr>
<td>School has sports yard</td>
<td>Argentina</td>
</tr>
<tr>
<td>School has choir</td>
<td>Colombia, Brasil</td>
</tr>
<tr>
<td>School has hall, club</td>
<td>Mexico, Peru</td>
</tr>
<tr>
<td>School has alumni association</td>
<td>Mexico</td>
</tr>
<tr>
<td>Percentage of students with minimum reading and writing materials</td>
<td>Mexico</td>
</tr>
<tr>
<td>School provides free meals</td>
<td>Mexico</td>
</tr>
<tr>
<td>Size of school library holdings</td>
<td>Argentina, Peru, Bolivia, Brazil</td>
</tr>
<tr>
<td>Can books be borrowed</td>
<td>Paraguay, Mexico, Bolivia</td>
</tr>
<tr>
<td>School has PTA</td>
<td>Colombia, Argentina</td>
</tr>
<tr>
<td>Admit students by residence</td>
<td>Colombia, Bolivia</td>
</tr>
<tr>
<td>Admit students by entrance exam</td>
<td>Colombia, Brazil</td>
</tr>
<tr>
<td>Group students by ability</td>
<td>Peru</td>
</tr>
<tr>
<td>Degree of principal emphasis on attendance</td>
<td>Mexico</td>
</tr>
<tr>
<td>Remedial programs</td>
<td>Mexico</td>
</tr>
<tr>
<td>Elective courses</td>
<td>Argentina</td>
</tr>
<tr>
<td>School requires students wear shoes</td>
<td>Peru, Bolivia</td>
</tr>
<tr>
<td>Afternoon teaching shift</td>
<td>Peru</td>
</tr>
<tr>
<td>Is teacher married</td>
<td>Argentina, Peru</td>
</tr>
<tr>
<td>Number of children teacher has</td>
<td>Colombia, Argentina</td>
</tr>
<tr>
<td>Number of persons living with teacher</td>
<td>Paraguay, Peru</td>
</tr>
<tr>
<td>Number of rooms in teacher’s house</td>
<td>Colombia</td>
</tr>
<tr>
<td>Teacher has car</td>
<td>Mexico</td>
</tr>
<tr>
<td>Teacher’s educational level</td>
<td>Bolivia, Brazil</td>
</tr>
<tr>
<td>Teacher specialized in education courses</td>
<td>Argentina, Bolivia, Peru</td>
</tr>
<tr>
<td>Attended teacher training courses</td>
<td>Mexico</td>
</tr>
<tr>
<td>Currently attending courses</td>
<td>Paraguay</td>
</tr>
</tbody>
</table>

**Note:** The following seven preschool variables were inserted in each country-specific regression prior to the block of school and teacher variables: father's education, mother's education, father's occupation, number of books in home, presence of phonograph, age of student, and age of student.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Countries in Which Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Educational level of teacher’s father</td>
<td>Peru, Bolivia, Paraguay, Argentina, Brazil</td>
</tr>
<tr>
<td>Occupational level of teacher’s father</td>
<td>Colombia</td>
</tr>
<tr>
<td>Number of schools ever taught in</td>
<td>Bolivia, Argentina, Colombia, Brazil</td>
</tr>
<tr>
<td>Working only at one school at a time</td>
<td>Paraguay</td>
</tr>
<tr>
<td>Permanent versus contract</td>
<td>Mexico</td>
</tr>
<tr>
<td>Spend time grading tests at school</td>
<td>Argentina, Colombia</td>
</tr>
<tr>
<td>Spend time at meetings</td>
<td>Mexico</td>
</tr>
<tr>
<td>Meet with parents</td>
<td>Paraguay</td>
</tr>
<tr>
<td>Gross monthly salary</td>
<td>Peru</td>
</tr>
<tr>
<td>Salary per hour of class</td>
<td>Bolivia</td>
</tr>
<tr>
<td>Satisfied with teaching profession</td>
<td>Peru, Mexico</td>
</tr>
<tr>
<td>Went into teaching for professional satisfaction</td>
<td>Mexico, Paraguay, Brazil</td>
</tr>
<tr>
<td>Believe opinion carries weight</td>
<td>Bolivia</td>
</tr>
<tr>
<td>Teacher has projector</td>
<td>Colombia, Paraguay</td>
</tr>
<tr>
<td>% of time explaining lesson</td>
<td>Peru, Brazil</td>
</tr>
<tr>
<td>% of time spent discussing exercises</td>
<td>Paraguay</td>
</tr>
<tr>
<td>% of time spent assigning homework</td>
<td>Argentina</td>
</tr>
<tr>
<td>% of time correcting exercises</td>
<td>Argentina</td>
</tr>
<tr>
<td>Books in teacher’s home</td>
<td>Bolivia</td>
</tr>
<tr>
<td>How often teacher reads periodicals</td>
<td>Mexico, Paraguay</td>
</tr>
<tr>
<td>How often teacher reads related to education</td>
<td>Paraguay</td>
</tr>
<tr>
<td>Annual cost of materials teacher asks student to buy</td>
<td>Bolivia</td>
</tr>
<tr>
<td>Teacher unwilling to label individual students as troublesome</td>
<td>Colombia</td>
</tr>
<tr>
<td>School requires uniform</td>
<td>Brazil</td>
</tr>
<tr>
<td>Number of class days per year</td>
<td>Brazil</td>
</tr>
<tr>
<td>Frequency teacher goes to library</td>
<td>Brazil</td>
</tr>
<tr>
<td>Teacher owns television</td>
<td>Brazil</td>
</tr>
<tr>
<td>Presence of science laboratory</td>
<td>Brazil</td>
</tr>
</tbody>
</table>

Note: The following seven preschool variables were inserted in each country-specific regression prior to the block of school and teacher variables: father’s education, mother’s education, father’s occupation, number of books in home, presence of phonograph, age of student, and age of student.

Source: S. P. Heyneman and W. A. Loxley "The Effect of Primary-School Quality on Academic Achievement Across Twenty-nine High and Low-Income Countries", The American Journal of Sociology, Vol 88, No.6 (May 1983).
Sample Questionnaire from Colombia’s Entrance Exams

**FORMULARIO DE INSCRIPCION**

**ADMISION A COLEGIOS OFICIALES**

**INSTRUCCIONES**

1. Conteste este formulario con la ayuda de su maestro.
2. No marque los óvalos que aparecen en la parte sombreada.
3. Rellene totalmente el óvalo correspondiente a cada pregunta.
4. Este formulario se debe llenar con LAPIZ de mina negra No. 2.

---

### I. DATOS PERSONALES

<table>
<thead>
<tr>
<th>PRIMER APELLIDO</th>
<th>SEGUNDO APELLIDO</th>
<th>NOMBRE(S)</th>
</tr>
</thead>
</table>

### II. DATOS EDUCATIVOS Y LABORALES

2. Estudió en algún jardín infantil, pre-escolar o kinder?  
   - 1 = SÍ  
   - 2 = NO

3. Jornada de Estudio de Educación Primaria  
   - 1 = MAÑANA  
   - 2 = TARDE

4. Trabaja?  
   - 1 = SÍ  
   - 2 = NO

5. Late noche?  
   - 1 = SÍ  
   - 2 = NO

### III. DATOS FAMILIARES (acerca de sus padres)

6. Vive?  
   - PADRE 1 = SÍ  
   - MADRE 1 = SÍ

7. Rinde con Usted?  
   - PADRE 1 = SÍ  
   - MADRE 1 = SÍ

### B. Nivel Educativo del Padre

- 1 = BÁSICA
- 2 = SECUNDARIA
- 3 = UNIVERSITARIA

### C. Nivel Educativo de la Madre

- 1 = BÁSICA
- 2 = SECUNDARIA
- 3 = UNIVERSITARIA

### D. Ocupación u oficio del Padre

1. DUEÑO DE GRAN EMPRESA, ADMINISTRADOR, GERENTE.  
2. PROFESIONAL, PILOTO U OFICIAL.  
3. EMPLEADO DE OFICINA, MAESTRO, SECRETARIO.  
4. PEQUEÑO COMERCiANTE, TRABAJADOR INDEPENDIENTE.  
5. CONDUCTOR, CARTERO, MENSajERO, OBRERO, POLICiA O SOLDADO.  
6. TRABAJADOR DEL CAMPO, LA MINERiA O LA PESCA.  
7. ARTESANO, TRABAJADOR DE LOS SERVICIOS, VENDEDOR AMBULANTE.  
8. DESEMPLEADO  
9. NO TRABAJA (JUBILADO, HOGAR, INVALIDO)

### E. Ocupación u oficio de la Madre

1. DUEÑO DE GRAN EMPRESA, ADMINISTRADOR, GERENTE.  
2. PROFESIONAL, PILOTO U OFICIAL.  
3. EMPLEADO DE OFICINA, MAESTRO, SECRETARIO.  
4. PEQUEÑO COMERCiANTE, TRABAJADOR INDEPENDIENTE.  
5. CONDUCTOR, CARTERO, MENSajERO, OBRERO, POLICiA O SOLDADO.  
6. TRABAJADOR DEL CAMPO, LA MINERiA O LA PESCA.  
7. ARTESANO, TRABAJADOR DE LOS SERVICIOS, VENDEDOR AMBULANTE.  
8. DESEMPLEADO  
9. NO TRABAJA (JUBILADO, HOGAR, INVALIDO)
**FORMULARIO DE INSCRIPCION**

RECUERDE QUE: Debe llenar este formulario SOLAMENTE con

Cuando encuentra dos o más alternativas para una respuesta, marque

por lo menos un circulo.

**I.- DATOS PERSONALES**

<table>
<thead>
<tr>
<th>NOMBRES Y APELLIDOS</th>
<th>DOCUMENTO DE IDENTIDAD</th>
<th>LUGAR DE NACIMIENTO</th>
<th>LUGAR DE RESIDENCIA ACTUAL</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**II.- INGRESOS Y SOSTENIMIENTO**

<table>
<thead>
<tr>
<th>SOSTENIMIENTO PERSONAL</th>
<th>INGRESO FAMILIAR</th>
<th>OTROS INGRESOS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**III.- INFORMACION SOBRE LOS PADRES**

<table>
<thead>
<tr>
<th>EDAD PADRES</th>
<th>NIVEL EDUCATIVO</th>
<th>OCCUPACIÓ U OFICO</th>
<th>CARACTERISTICAS DEL EMP.</th>
<th>DEPENDENCIAS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**IV.- OTRAS ACTIVIDADES**

<table>
<thead>
<tr>
<th>EXPERIENCIA DE TRABAJO</th>
<th>CARACTERISTICAS DEL EMP.</th>
<th>DEPENDENCIAS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**V.- EDUCACION MEDIA**

<table>
<thead>
<tr>
<th>CALENDARIO</th>
<th>HORARIO</th>
<th>NATURALIZA DEL PLANTE</th>
<th>CALENDARIO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**VI.- DEPORTES**

<table>
<thead>
<tr>
<th>DEPORTE</th>
<th>NIVEL</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**ANNEX 4**

Page 2
INFORME DE LOGRO DE OBJETIVOS ACADEMICOS 1989

RESULTADOS NACIONALES
ASIGNATURA: CIENCIAS NATURALES

<table>
<thead>
<tr>
<th>OBJETIVOS ESPECIFICOS DE 8º AÑO BASICO</th>
<th>ALUMNOS QUE RINDIERON LA PRUEBA:</th>
<th>Promedio establecimientos municipales</th>
<th>Promedio establecimientos subvencionados</th>
<th>Promedio establecimientos priv. pagados</th>
<th>Promedio nacional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conocer y comprender que las poblaciones de seres vivos requieren de ciertas condiciones básicas para subsistir.</td>
<td>13.428</td>
<td>69.5</td>
<td>71.0</td>
<td>88.3</td>
<td>71.4</td>
</tr>
<tr>
<td>Comprender que las relaciones de interdependencia entre los seres vivos y su medio son factores que favorecen su sobrevivencia.</td>
<td>4.948</td>
<td>54.5</td>
<td>53.0</td>
<td>71.9</td>
<td>55.6</td>
</tr>
<tr>
<td>Comprender que los seres vivos interactúan entre sí y con su ambiente, generando equilibrio natural.</td>
<td>1.695</td>
<td>58.0</td>
<td>57.5</td>
<td>76.7</td>
<td>59.4</td>
</tr>
<tr>
<td>Comprender la importancia de la conservación del medio ambiente y el rol del proceso reproductivo en la supervivencia de la organización biológica.</td>
<td>20.071</td>
<td>53.9</td>
<td>56.0</td>
<td>77.9</td>
<td>56.4</td>
</tr>
<tr>
<td>Comprender que el organismo humano requiere un constante aporte de alimento, oxígeno, y de la eliminación de productos de desecho. Comprender la importancia de la conservación y mejoramiento de la salud.</td>
<td>67.3</td>
<td>68.9</td>
<td>79.9</td>
<td>68.8</td>
<td></td>
</tr>
<tr>
<td>Comprender que el ser humano requiere de sistemas de relaciones para sobrevivir en un medio cambiante.</td>
<td>61.7</td>
<td>60.5</td>
<td>83.3</td>
<td>63.2</td>
<td></td>
</tr>
<tr>
<td>Comprender que el ser humano se perpetúa a través del proceso de la reproducción y valorar su dimensión eco-social.</td>
<td>61.3</td>
<td>77.0</td>
<td>61.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conocer algunas características y propiedades de la materia, los cambios que puede experimentar, y otras manifestaciones de la energía y sus transformaciones.</td>
<td>45.8</td>
<td>48.9</td>
<td>66.8</td>
<td>48.4</td>
<td></td>
</tr>
<tr>
<td>Comprender que la materia y la energía se transforman y que en la naturaleza la totalidad de ambas se conserva.</td>
<td>48.5</td>
<td>50.5</td>
<td>69.4</td>
<td>50.8</td>
<td></td>
</tr>
<tr>
<td>Comprender que los cambios de posición de un objeto o ser vivo están en función del espacio y del tiempo.</td>
<td>56.3</td>
<td>60.0</td>
<td>82.5</td>
<td>59.4</td>
<td></td>
</tr>
<tr>
<td>Conocer las características fundamentales del Sistema Solar.</td>
<td>6.7</td>
<td>50.5</td>
<td>71.9</td>
<td>51.1</td>
<td></td>
</tr>
<tr>
<td>Conocer las variables que influyen en el tiempo atmosférico y caracterizar los climas como el conjunto de registros meteorológicos a lo largo del tiempo.</td>
<td>47.1</td>
<td>51.2</td>
<td>78.7</td>
<td>50.8</td>
<td></td>
</tr>
<tr>
<td>Conocer los agentes geológicos a los cuales provocan cambios en el suelo.</td>
<td>59.8</td>
<td>60.5</td>
<td>77.7</td>
<td>61.5</td>
<td></td>
</tr>
</tbody>
</table>

REQUISITOS GENERALES

| | 56.5 | 57.9 | 77.2 | 58.6 |
1. Dominio de Conocimientos (Objetivos) Medidos en la Prueba
(Cada objetivo fue medido por 3 ítems).
Se Domina un Objetivo cuando el Promedio es mayor o igual a 2.00

<table>
<thead>
<tr>
<th>Número de Objetivo</th>
<th>Puntaje Promedio por Objetivo en su Colegio</th>
<th>Puntaje Promedio por Objetivo en el Circuito</th>
<th>Puntaje Promedio por Objetivo en la Región</th>
<th>Puntaje Promedio por Objetivo Nacional</th>
<th>% de Estudio que domina el Objetivo en su Colegio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.17</td>
<td>1.83</td>
<td>1.93</td>
<td>2.12</td>
<td>44.44</td>
</tr>
<tr>
<td>2</td>
<td>0.71</td>
<td>1.02</td>
<td>0.82</td>
<td>1.05</td>
<td>23.50</td>
</tr>
<tr>
<td>3</td>
<td>0.49</td>
<td>0.89</td>
<td>0.91</td>
<td>1.04</td>
<td>14.29</td>
</tr>
<tr>
<td>4</td>
<td>0.82</td>
<td>0.64</td>
<td>2.71</td>
<td>0.97</td>
<td>17.65</td>
</tr>
<tr>
<td>5</td>
<td>0.53</td>
<td>1.76</td>
<td>2.33</td>
<td>2.69</td>
<td>17.65</td>
</tr>
<tr>
<td>6</td>
<td>1.63</td>
<td>2.02</td>
<td>2.03</td>
<td>2.39</td>
<td>57.14</td>
</tr>
<tr>
<td>7</td>
<td>1.14</td>
<td>1.64</td>
<td>1.70</td>
<td>2.09</td>
<td>37.14</td>
</tr>
<tr>
<td>8</td>
<td>1.22</td>
<td>1.33</td>
<td>1.70</td>
<td>1.75</td>
<td>44.40</td>
</tr>
<tr>
<td>9</td>
<td>1.29</td>
<td>1.52</td>
<td>1.70</td>
<td>1.73</td>
<td>37.14</td>
</tr>
<tr>
<td>10</td>
<td>0.35</td>
<td>0.67</td>
<td>1.13</td>
<td>1.25</td>
<td>6.00</td>
</tr>
<tr>
<td>11</td>
<td>1.11</td>
<td>0.95</td>
<td>1.39</td>
<td>1.43</td>
<td>18.27</td>
</tr>
<tr>
<td>12</td>
<td>0.28</td>
<td>0.38</td>
<td>0.79</td>
<td>1.00</td>
<td>3.61</td>
</tr>
<tr>
<td>13</td>
<td>0.74</td>
<td>0.87</td>
<td>0.98</td>
<td>1.02</td>
<td>22.36</td>
</tr>
<tr>
<td>14</td>
<td>0.47</td>
<td>0.79</td>
<td>1.17</td>
<td>1.14</td>
<td>5.08</td>
</tr>
<tr>
<td>15</td>
<td>0.69</td>
<td>0.92</td>
<td>1.13</td>
<td>1.22</td>
<td>11.43</td>
</tr>
<tr>
<td>16</td>
<td>0.56</td>
<td>0.88</td>
<td>1.01</td>
<td>1.17</td>
<td>5.58</td>
</tr>
<tr>
<td>17</td>
<td>0.88</td>
<td>1.07</td>
<td>1.24</td>
<td>1.35</td>
<td>23.53</td>
</tr>
<tr>
<td>18</td>
<td>1.4</td>
<td>1.81</td>
<td>1.74</td>
<td>1.90</td>
<td>34.25</td>
</tr>
<tr>
<td>19</td>
<td>0.59</td>
<td>0.67</td>
<td>0.95</td>
<td>1.05</td>
<td>11.75</td>
</tr>
<tr>
<td>20</td>
<td>1.78</td>
<td>1.69</td>
<td>1.95</td>
<td>2.05</td>
<td>61.11</td>
</tr>
<tr>
<td>21</td>
<td>1.33</td>
<td>1.31</td>
<td>1.68</td>
<td>1.62</td>
<td>44.44</td>
</tr>
<tr>
<td>22</td>
<td>0.63</td>
<td>0.73</td>
<td>0.77</td>
<td>0.85</td>
<td>6.57</td>
</tr>
<tr>
<td>23</td>
<td>0.94</td>
<td>0.93</td>
<td>1.00</td>
<td>1.01</td>
<td>11.43</td>
</tr>
</tbody>
</table>
Ministerio de Educación Pública
Instituto de Investigación para el
Mejoramiento de la Educación Costarricense (IIMEC)

Ejemplo de Informe Enviado a Escuelas y Colegios
1988

Paso Candas
Resultados Obtenidos en Sexto Grado
Matemática

1. Calificación Obtenida en la Prueba

<table>
<thead>
<tr>
<th>Categoría de Nota</th>
<th>Número de Estudiantes</th>
<th>Porcentajes Estudiantes</th>
</tr>
</thead>
<tbody>
<tr>
<td>00-20</td>
<td>3</td>
<td>8.57</td>
</tr>
<tr>
<td>21-40</td>
<td>28</td>
<td>80.00</td>
</tr>
<tr>
<td>41-59</td>
<td>4</td>
<td>11.43</td>
</tr>
<tr>
<td>60-70</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>71-80</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>81-100</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

2. Resumen de Notas Promedio
   Si aparece un "S" este resultado fue obtenido con
   Menos de seis estudiantes

   Nota Promedio de su Institución 31.01

NOTA PROMEDIO DE LAS ESCUELAS A NIVEL GENERAL
(INCLUYE TODOS LOS TIPOS DE DIRECCION)

Nota Promedio Nacional 49.11
Nota Promedio de su Región 43.71
Nota Promedio de su Circuito 39.24
Nota Promedio de la mejor Escuela a Nivel Nacional 93.75
Nota Promedio de la Mejor Escuela de su Región 79.86

NOTA PROMEDIO DE LAS ESCUELAS TIPO - DEG83

Nota Promedio Nacional 47.52
Nota Promedio de su Región 39
Nota Promedio de su Circuito 31.01
Nota Promedio de la Mejor Escuela a Nivel Nacional 72.46
Nota Promedio de la Mejor Escuela de su Región 50.37
GUIA DE INFORMACION
SOBRE LAS PRUEBAS
DE CONOCIMIENTOS DE
EDUCACION BASICA PRIMARIA

Dirigida a los docentes de 5o. de Primaria

OBJETIVO DE LA PRESENTE INFORMACION:
- Ofrecer a los docentes información sobre el procedimiento seguido en la elaboración de las pruebas.
- Ofrecer información sobre los contenidos y ejemplos de las pruebas con el fin de familiarizar a los docentes y sus alumnos, con el manejo de las mismas.

PROCEDIMIENTO SEGUIDO EN LA ELABORACION DE LAS PRUEBAS

De acuerdo con la resolución 14584/Agosto 1979, el Servicio Nacional de Pruebas del ICFES tiene a su cargo el programa de Admisión a Colegios Oficiales de Bogotá, D.E. Este programa exige la elaboración de pruebas objetivas de aptitud y conocimientos para el nivel de 5o. Primaria a los niños inscritos en el programa.

De acuerdo con el plan de desarrollo de la renovación curricular en el Distrito, para 1984 era necesario elaborar pruebas de conocimientos basadas en los nuevos programas y de acuerdo con el nuevo enfoque y metodología previstos en la renovación. Con este fin la Sección de Diseño de Pruebas del Servicio Nacional de Pruebas programó las siguientes actividades:
1. Taller para el planeamiento de las pruebas de conocimientos: Matemáticas, Ciencias Naturales, Español y Literatura y Ciencias Sociales.

El plan de trabajo de este taller incluyó:

- Información sobre la renovación curricular en general y de los programas en particular.
- Información sobre el plan de desarrollo del nuevo currículo trazado por la Secretaría de Educación de Bogotá, en las diferentes materias.
- Estudio detallado de los programas por cursos, enfatizando en los de 4o. y 5o.
- Determinación de los objetivos a evaluar y de los procesos correspondientes a ellos.
- Determinación de los contenidos correspondientes.
- Definición de la estructura de la prueba y sus especificaciones.
- Elaboración de la información para los docentes de 5o. de Educación Básica Primaria.

Este plan se complementó con sesiones de trabajo posteriores.

Participantes:

- División de Currículo Formal del MEN
- Dirección de Investigación para la Educación (DIE - CEP) de la Secretaría de Educación de Bogotá D.E.
- Docentes seleccionados por la DIE, en cada una de las materias.
- Sección de Diseño de Pruebas del SNP del ICFES.

2. Taller para la elaboración de las preguntas que conforman las pruebas, de acuerdo con las estructuras y especificaciones definidas en la etapa anterior.

Participantes:

Docentes de 5o. de Educación Básica Primaria en grupos por materia, dirigidos y coordinados por un psicólogo de la Sección de Diseño de Pruebas del SNP.

Revisión de la prueba por parte de los técnicos del MEN y DIE, quienes intervinieron en la etapa de planeamiento.

3. Presentación a la dirección del SNP del trabajo desarrollado.


---CONTINUE EN EL SIGUIENTE PLIEGUE (arriba)---
Entrega de la presente información a los supervisores, directores y docentes de 5o. de Educación Básica Primaria de las escuelas de Bogotá, D.E.


Estudio de resultados.

Es necesario recalcar que en la elaboración de esta información intervinieron los técnicos del MEN, de la DIE, docentes de 5o. de Primaria y técnicos del SNP, quienes desde un comienzo han venido participando en la construcción de estas pruebas y seguirán participando en su perfeccionamiento.

Igualmente les solicitamos a todos los docentes utilizar esta información y hacerlos llegar sus opiniones y sugerencias, ya que sabemos que son ustedes quienes con su interés y empeño pueden consolidar en el aula los propósitos de la renovación curricular. Mediante estas pruebas se pretende lograr una evaluación sumativa, que proporcione indicadores del rendimiento académico, apenas como puntos de referencia necesarios en el análisis de los múltiples aspectos que interactúan en el proceso de formación de los estudiantes.

PRUEBA DE MATEMATICAS

La prueba sigue el enfoque del “nuevo currículo”, basado en la teoría de sistemas, su metodología en la teoría de Piaget y se ciñe estrictamente a los contenidos de los programas. La prueba tiene 30 preguntas. Estas constan, en todos los casos, de un enunciado y de cuatro posibilidades de respuesta, de las cuales sólo una es correcta y se denomina clave; las opciones que no son correctas se llaman distractores.

Los grandes temas y su importancia en la prueba coinciden con los de los programas, ellos son:

- Sistemas numéricos
  - números naturales
  - números fraccionarios
  - números decimales
  - razones y proporciones
- Sistemas geométricos
- Sistemas métricos
- Sistemas de datos
- Sistemas lógicos
- Conjuntos
- Relaciones y operaciones

La prueba comprende preguntas que implican:

- Simplemente el recuerdo del material previamente enseñado o la realización de operaciones.

EJEMPLO 1

Objetivo: Aplicar en el cálculo numérico las propiedades de la adición y la multiplicación.

CONTINUEABAJO
Tema: Números naturales
Propiedad distributiva de la multiplicación respecto de la suma.

Pregunta: Cuál de los siguientes procedimientos NO conduce al resultado de efectuar: \(5 \times (4 + 3 - 2)\)

A- \((5 \times 4) + 3 - 2\)

B- \((5 \times 2) + (5 \times 3)\)

C- \((5 \times 4) + (5 \times 1)\)

D- \((5 \times 4) + (5 \times 3) - (5 \times 2)\)

Clave: A. Porque cualquiera de los otros procedimientos lleva al resultado correcto.

La transformación de los contenidos de una descripción verbal a una representación gráfica o viceversa, así como la comprensión de su significado.

EJEMPLO 2

Objetivo específico: Interpretar datos presentados en diferentes formas.

Tema: Sistemas de datos
Organización y representación gráfica

Pregunta: La producción ganadera del país en un período de tiempo, por regiones naturales es:

<table>
<thead>
<tr>
<th>Región Natural</th>
<th>Caribe</th>
<th>Pacífico</th>
<th>Amazonia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cabezas de ganado</td>
<td>18.000</td>
<td>7.000</td>
<td>8.000</td>
</tr>
</tbody>
</table>

Cuál de los siguientes cuadros la representa mejor?

Clave: B. Porque es la única que muestra la coincidencia entre los componentes de los ejes “x” y “y”, de acuerdo con los datos dados.
EJEMPLO 3

Objetivo específico: Reconocer relaciones de orden multiplicativo.

Tema: Relaciones y operaciones.
Relación de orden

Pregunta: Cuál de las siguientes series NO obedece la misma regla?
A. 6, 12, 24, ...  
B. 1, 2, 4, ...  
C. 4, 8, 12, ...  
D. 7, 14, 28, ...

Clave: C. Porque es la única que no satisface la relación común a las demás: "... es el doble de ..."

- Aplicar los conocimientos en situaciones concretas, o sea, resolver problemas.

EJEMPLO 4

Objetivo específico: Resolución e interpretación de problemas.

Tema: Razones y proporciones.
Proporcionalidad directa

Pregunta: Un panadero hace panes de $20,00 cuatro veces por semana:
El martes con 500 gr obtiene 30 panes
El jueves con 750 gr obtiene 45 panes
El sábado con 1.250 gr obtiene 65 panes
El domingo con 250 gr obtiene 15 panes

¿Qué día reclamarán los clientes porque los panes son más pequeños?
A. el martes  
B. el jueves  
C. el sábado  
D. el domingo

Clave: C. Porque al comparar las magnitudes, directamente proporcionales, es la única que se encuentra con diferente razón (cociente).
Other Reports in the Series


No. 3: Easing the Poor Through Economic Crisis and Adjustment: The Story of Bolivia's Emergency Social Fund, May 1991

No. 4: Direct Credit for Privatized Firms, June 1991


No. 6: Mexico Labor Retraining Program: Poverty Alleviation and Contribution to Growth, August 1991


No. 8: Choice of Nutritional Status Indicators for Young Children in Public Health Programs, September 1991