Multi-dimensional evaluation: measures for both right and left sides of the equation

Marlaine E. LOCKHEED
Department Population and Human Resources
World Bank, Washington

The evaluation of primary and secondary schools.
The autonomy of institutions and quality control
Abstract

Monitoring educational progress requires data on both student learning and determinants of achievement. It also requires measures of learning achievement that can provide information on trends over time. This paper (a) reviews basic concepts related to measuring learning, educational inputs and processes, (b) compares assessments to monitor trends with measures of achievement for other purposes and (c) estimates costs for different purposes of testing.
The theme of this conference is "ensuring the success of all in the basic school." But what is meant by success in the basic school? What must schools do to ensure the success of their students? And what tools enable a nation to monitor progress toward success? There are many definitions of success in the basic school, although most educators agree that basic education should provide children with basic literacy, numeracy and reasoning skills. These learning objectives, however, are established by national or local education authorities, and for this session I will assume that success means success with respect to some learning objective; this is what I call the left-hand side of the equation.\footnote{Educational learning objectives are not limited to cognitive skills. For example, Lockheed and Vepsal open note that "Primary education serves several purposes. One objective is to teach students basic cognitive skills. A second aim is to develop attitudes and skills necessary to function effectively in society. Third, primary education contributes to nation-building objectives." (Lockheed and Vepsal, 1991:vii)}

On the right-hand side of the equation, there are three types of determinants of learning achievement: what students bring to school with them from their families and communities, what schools and classrooms provide students in terms of material and non-material inputs, and how these inputs are used in terms of school and classroom processes.
Many research studies in developed and developing countries have collected data on both sides of the equation at one point in time. These cross-sectional studies provide a snapshot of the condition of education at that time. For example, the ECIEL studies in six countries in Latin America in the 1970s, the IEA studies in up to 40 countries since the 1960s, IREDU studies currently ongoing in three Francophone African countries, ongoing studies by the IIIEP in six countries, and Unesco studies in several countries, to name but a few. These studies represent an important step forward. But more is needed. None of these efforts have addressed the very significant problem of how to construct measures of learning achievement on the left-hand side that enable a country to monitor progress in reaching its learning goals over time. Moreover, only a few of these studies have been designed to inform national decision making in education, and provide information on how to reach these learning goals.

Good management requires the systematic collection and use of information that support efforts to improve school effectiveness. Information must be useful and made available to policymakers whose decisions affect the delivery of educational services. An essential tool of monitoring is a data-gathering and analysis system that includes measures of learning (the left-hand side of the equation), measures or indicators of inputs and processes (the right-hand side of the equation), the ability to monitor trends over time, and analyses that draw inferences for action from the entire equation. This presentation will address these four issues as they relate to developing countries.
1. Measuring learning: The left-hand side of the equation.

Tools to measure student learning routinely are essential for improving the quality of education. These tools are achievement tests. Achievement tests are defined by three elements: their content, their behavioral objectives and their format. The content objective of a test is the subject matter that is being tested, such as vocabulary, historical events, or long division. The content objective of a test is largely determined by the curriculum (although in some countries, test developers and curriculum specialists do not keep close company).

The "behavioral objective" of a test is the type of skill being tested. In 1956, Benjamin Bloom taxonomized cognitive "behavioral objectives" into six elements: knowledge, comprehension, application, analysis, synthesis, and evaluation. Bloom's taxonomy has been widely used over the past 35 years to characterize both curricula and tests. Testing knowledge typically requires the student to recall the meaning of terms and specific facts. Testing the "higher order" skills of comprehension, application, analysis, synthesis and evaluation requires the student to demonstrate that he or she understands the material, can use information in a concrete situation, break down material into its parts, can assemble material into a whole, and can judge the value of a thing for a given purpose using definite criteria (Bloom 1956). Most test in developing countries concentrate on the recall of facts.
Test format refers to the physical design of the items on the test. Test items can be developed in a wide variety of formats, which are often classified into "objective" test formats and "performance" test formats. Objective test items can be written in several different formats, not only the "multiple-choice" items often employed in mass testing. Some other types of item formats are: matching, question and short answer, rank-ordering, statement and comment (short essay). Objective test formats are comparatively expensive to develop but inexpensive to score.

Performance tests, which are becoming increasingly the format of choice among educational theorists, require the student to demonstrate knowledge and skills through an actual or simulated performance. Thus, a student may submit a videotape of a dance performance, a portfolio of his or her works of art; he or she may be asked to carry out a short science experiment to demonstrate scientific understanding, to write an essay on a specific topic to demonstrate writing skills, or to repair an automobile engine. The closer the performance is to reality, the more "authentic" it is viewed to be. Performance tests range widely with respect to the cost of development, but are uniformly expensive to score reliably.

2. Measures of Indicators of inputs and processes: The right-hand side of the equation

To complement achievement measures, information on the characteristics of students, and the costs and characteristics of classrooms and schools are required. This information
can help clarify the meaning of the achievement test scores, particularly if they are used for purposes of monitoring schools or teacher effectiveness. For example, the average achievement test scores of schools are often reported publicly, and are cited as evidence of the quality of the school's instructional program. Yet it is well known that the composition of the school's student body, with respect to students' socioeconomic background or prior level of achievement, is an important determinant of the school's average achievement. Good schools are those that boost student achievement, not those that have the good fortune to recruit good students.

To take into account differences between schools in their student intake, and in the resources and process available in classrooms and schools requires the collection of additional data. Many researchers have sought to identify the minimum set of information that should be collected. Most recommend the collection of basic intellectual, socioeconomic and cultural background data on students. Examples include the students' sex, age, nutritional status, socio-economic status, language, attitudes, expectations, and ability. Theory, research and experience suggest four classroom inputs that are important: the curriculum, instructional materials, adequate time for learning, and effective teaching (Lockheed and Verspoor 1991). Beyond these inputs are managerial processes at the schools and within the classroom, such as tracking, school structure and autonomy and school context conditions such as student access to knowledge, organizational press for achievement, and professional teaching conditions (Oakes 1989).
3. Monitoring trends over time

To track progress requires the ability to monitor trends. Monitoring trends over time on the right-hand side of the equation is relatively simple. Standardized measures exist. Counts of textbooks, minutes of instructional time, per-student expenditures, categories of teacher training all use measures or categories -- numbers, minutes, dollars -- that do not change their meaning from one time to the next. For example, an evaluation of a massive rural education development program in northeast Brazil (Harbison and Hanushek, forthcoming) monitored progress in such "right-side" variables as the percentage of teachers with more than primary-level schooling (38%, 41% and 47% in 1981, 1983 and 1985 respectively), the percentage of schools with two or more classrooms (52%, 66% and 74% in 1981, 1983 and 1985 respectively), and the percentage of schools receiving grade 1 textbooks (81% in 1885 versus 30% in 1983).

However, monitoring trends over time on the left side of the equation -- student learning -- is more difficult because most existing measures of learning in developing countries do not refer to an inelastic standard, such as minutes or counts. Instead, the standard can and does change in meaning from one year to the next. Yet, to monitor trends over time, measures need to be invariant, and achievement tests in developing countries need to be designed accordingly. This is possible. No matter what the content objective, behavioral objective or format of the test, it is possible to design a test that can monitor achievement over time.
For test scores to be meaningful over time, student performance must be measured against an inelastic yardstick of achievement, such as a norm reference group or criteria references (Anastasi 1988; Berk 1984). In addition, they must be "standardized" with respect to content and behavioral objectives, format, administration procedures and scoring. Standardizing the content requires that the same or equivalent questions or performance tasks be posed for all students. Standardizing test administration requires uniformity in the written and verbal instructions given to examinees, in the length of time afforded them, in the materials provided to them, and in the physical testing environment. Standardized scoring requires explicit, impartial procedures for correcting tests or judging performance. One of the biggest obstacles to monitoring student learning over time in developing countries is the lack of such tests.

This is not to say that countries are inexperienced with testing. They are not. But most countries are experienced with two types of unstandardized tests: (a) those used for selection or certification, which are created new each year only to certify the academic accomplishment or to select for secondary or higher education students in a single age or grade cohort (examinations), and (b) those used by teachers to assess the performance of individual students over the course of the year (continuous student assessment).

Examinations. One survey of assessment systems in 17 countries reported that 80% of the tests were used for certification or selection (Binkley, Guthrie and Wyatt 1991). A recent World Bank survey of five Anglophone and 14 Francophone African countries found
virtually all tests were designed for selection and certification purposes (Greaney and Kellahan 1991). Selection examinations are useful in rationing scarce places and can influence instruction. But they are not designed to provide meaningful measures of accomplishment from one year to the next. Instead, because of the need to maintain test security, the items on the tests are completely different from one year to the next. This has implications for the extent to which they can make reference to an "inelastic yardstick of performance." For example, it means that a norm-reference group, which was administered exactly the same items at some earlier time period, cannot be employed to provide the standard. It also means that performance or criterion-referenced selection or certification tests may be unable to provide a measure of change over time.

While many countries attempt to create tests that are of similar difficulty each year, most countries do not test this equivalence empirically, because standard psychometric techniques rely to some extent upon the reuse of at least some items. Instead, countries must rely on subjective judgment. These judgments are often less accurate than empirically based equating procedures. As a result, wide fluctuations in test scores are often reported. For example, in one Anglophone African country, the proportion of students who passed the Grade 7 mathematics test dropped from 83% in one year to 55% the following year -- undoubtedly due to differences in the difficulty of the test, not to differences in what cohorts of students only one year apart actually learned. Because of their strong security needs, selection and certification tests are not likely to be standardized with respect to content and
behavioral objective, and are often not standardized in other respects as well; they are therefore not useful for monitoring trends.

*Continuous student assessment.* Regular monitoring of student performance is an essential teaching tool. It can be accomplished in two ways. First, the teacher can regularly assess student performance through such informal or unstandardized devices as quizzes, written assignments and projects. Second, the teacher can administer to students formally developed tests, such as are often prepared by testing agencies and publishers. In most developing countries, formally developed tests for monitoring individual student progress are non-existent, and the only type available are those developed by teachers. Because they are completely unstandardized, they are also unable to provide useful data for monitoring trends.

4. Putting it all together: Assessments for educational management

Educational managers need essentially three types of information: (a) general information about how well the education system is doing with respect to achieving its goals, (b) specific information about the performance of individual schools or groups of schools (for example, school districts or geographic regions), and (c) analyses of the implementation and effects of policies adopted for school improvement. General information provides the global picture, school or regional information helps managers hold schools or regions accountable for their performance, and the analytic work help to guide future investment
decisions. How well can existing types of examinations and tests provide these types of information?

*Characteristics of tests.* Table 1 contrasts the characteristics of selection/certification tests and tests to monitor individual students' performance with the characteristics of tests for monitoring progress, school accountability and evaluation. The tests for these five purposes are, or can be, similar in terms of their content and behavioral objectives and their format, although selection tests are likely to be more difficult than tests for management purposes. They are significantly different, however, in terms of the number of students tested, the degree to which the tests are standardized, their use of external reference criteria and the extent of supplementary measures collected. Tests for management purposes can be administered to samples of students; tests for selection, certification or individual performance assessment need to be administered to all students. Tests for management purposes need to be standardized with respect to their content, administration and scoring; tests for selection are frequently not standardized for either administration or scoring (and sometimes not even for content). Tests for monitoring progress must make reference to an external criteria; other types of test often do not. Tests to evaluate the effectiveness or efficiency of specific educational policies must include measures of these policies, as well as measures of other student and school characteristics; tests for other purposes use supplementary measures only to a limited extent.
<table>
<thead>
<tr>
<th>Purpose</th>
<th>Monitoring Progress Toward National Educational Goals</th>
<th>Evaluating Effectiveness and Efficiency of Specific Policies</th>
<th>Holding Schools Accountable for Performance</th>
<th>Selecting and/or Certifying Students</th>
<th>Teacher Assessment of Individual Student Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example</td>
<td>National Assessment of Educational Progress, USA</td>
<td>Evaluation Analytique de l’enseignement Primaire, Benin</td>
<td>National Grade 7 Evaluation Thailand</td>
<td>Certificate of Primary Education, Kenya</td>
<td>National Assessment, Britain and Wales</td>
</tr>
<tr>
<td>To Whom Administered</td>
<td>Sample of students in selected age or grade cohorts</td>
<td>Sample of students in sample of schools</td>
<td>Sample of students in sample of schools</td>
<td>All students in terminal year</td>
<td>All students, in selected age cohorts</td>
</tr>
<tr>
<td>When Administered</td>
<td>Periodically (annual, biannual)</td>
<td>One time</td>
<td>One time</td>
<td>Annual</td>
<td>Annual</td>
</tr>
<tr>
<td>Classroom Objective</td>
<td>Selected domains (e.g., mathematics, science)</td>
<td>Selected domains</td>
<td>Selected domains</td>
<td>All domains of curriculum</td>
<td>Selected domains</td>
</tr>
<tr>
<td>Behavioral Objective</td>
<td>Knowledge and higher order thinking skills</td>
<td>Knowledge and higher order thinking skills</td>
<td>Knowledge and higher order thinking skills</td>
<td>Knowledge and higher order thinking skills</td>
<td>Knowledge and higher order thinking skills</td>
</tr>
<tr>
<td>Normal Objective Performance</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Standardized</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Calculated</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Administration</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Scoring</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Reference</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Reference criteria</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Supplementary measures</td>
<td>Limited</td>
<td>Yes</td>
<td>Limited</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Student background</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Classroom/school inputs</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Classroom/school procedures</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
Costs. Total recurrent costs for tests are determined largely by the numbers of students that are tested and the type of test items used -- objective items or performance items. To monitor progress, it is possible to obtain reliable measures with scientifically drawn samples of as few as 2000 students. Evaluations often use a small sample of schools (say, 1%) and sample students within schools. School performance accountability requires testing samples of students at all schools. Selection and certification examinations typically test all students in all schools, during the students' terminal year of study. Teacher assessments involve all students in all grades. Thus, assessments that monitor progress and studies that evaluate policies involve only relatively few students and schools, while school accountability schemes, selection and certification examinations and teacher assessment of student performance involve much larger numbers of schools and students.

If, for discussion, we assume that objective tests cost about $1 to produce and score, and performance tests cost $10 to produce and score, then it is possible to estimate what it would cost to carry out different types of tests in average low-income (excluding China and India), lower-middle income and upper-middle income countries. Table 2 presents these estimates, which are based on the average number of students and schools in countries at three levels of economic development. While these figures are hypothetical, they give an idea of the order of magnitude involved.
Recurrent costs associated with teacher assessment of student progress are the highest, assuming that all primary students are tested annually and that the tests are neither produced nor scored by teachers are part of their regular work. In the average developing country (other than India and China), testing all students with formally developed tests would cost between $1.6 and $4 million to produce and score tests having objective items and between $15 and $41 million to produce and score tests having performance items. Tests for selection and/or certification are next expensive, amounting to up to $700,000 for objective tests and up to nearly $7 million for performance tests in the average developing country. Tests for school accountability are approximately the same cost as those for selection and certification in low income countries, but only about half the cost in middle income countries; this is because higher proportions of children reach a terminal primary school grade in middle income countries than in low income countries. Tests for monitoring the progress of the education system or for evaluating education policies are by far the lowest cost activities, costing less than one-tenth of one percent of the cost of testing all students in all grades. By using samples of students to monitor national progress, more expensive test formats could be employed. Developing countries could afford to use performance tests for monitoring progress or in evaluation exercises and it would cost less than two-tenth of one percent of the cost of administering and scoring objective tests for selection or certification.

Thus, for many developing countries, it may be infeasible to test all children annually. Sample based surveys and specially designed standardized tests for monitoring progress are
<table>
<thead>
<tr>
<th>Purpose</th>
<th>Teacher selection and/or certifying of individual student performance</th>
<th>Average of primary school/county (n=800)</th>
<th>Objective tests</th>
<th>Performance tests</th>
<th>School and/or district population (n=800)</th>
<th>School and/or district population (n=800)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monitoring program toward national educational goals</td>
<td>All schools; all students in terminal grade</td>
<td>2,000%</td>
<td>1/6 sample of schools; 20% of students per school</td>
<td>All schools; all students in terminal grade</td>
<td>$1,578,004</td>
<td></td>
</tr>
<tr>
<td>Evaluating effectiveness of specific policies</td>
<td>All schools; all students in terminal grade</td>
<td>2,000%</td>
<td>1/6 sample of schools; 20% of students per school</td>
<td>All schools; all students in terminal grade</td>
<td>$1,578,004</td>
<td></td>
</tr>
<tr>
<td>Holding accountable for performance</td>
<td>All schools; all students in terminal grade</td>
<td>2,000%</td>
<td>1/6 sample of schools; 20% of students per school</td>
<td>All schools; all students in terminal grade</td>
<td>$1,578,004</td>
<td></td>
</tr>
</tbody>
</table>
capable of providing nearly all the same information to a manager as tests administered to all students, but at a small fraction of the recurrent cost. In such a combined tool, the left-hand side of the equation uses standardized objective or performance tests that are equated for content objective and skill objectives over time. On the right-hand side of the equation, sufficient measures of student, classroom and school characteristics are collected to enable managers and policy makers to evaluate the implementation of policy changes, assess their effects on achievement, and estimate the cost-effectiveness of alternative investment choices. Instead of testing whole populations of students, samples of students are tested. Results are reported publicly and rapidly. Analyses of student responses to items inform teachers and the public about content objectives and behavioral objectives that remain to be met. Questions of efficiency can be addressed by combining measures of learning achievement with cost information (Lockheed and Hanushek 1988). Such a system is beneficial in five ways:

1. *It enables longitudinal comparisons through the adoption of standardized testing procedures.*

2. *It provides a standard measure for system accountability.*

3. *It can improve pedagogy.*

4. *Scientific sampling lowers the total cost associated with testing.*
5. It permits rapid monitoring of the implementation, impact on achievement, and cost-effectiveness of policies and programs.
References


