PREAL
Programa de Promoción de la Reforma Educativa en América Latina y el Caribe
Partnership for Educational Revitalization in the Americas

Working Paper Series

No. 40

The Educational Assessments That Latin America Needs

by Pedro Ravela, Patricia Arregui, Gilbert Valverde, Richard Wolfe, Guillermo Ferrer, Felipe Martínez Rizo, Mariana Aylwin, and Laurence Wolff

March 2008

The authors are members of the Steering Committee of the PREAL Working Group on Standards and Assessments (GTEE). The opinions expressed in this study are those of the authors and do not necessarily reflect the positions of PREAL or its sponsoring institutions.

For additional copies, contact PREAL at the Inter-American Dialogue. This report can be downloaded from www.preal.org/publicacion.asp.

Inter-American Dialogue
1211 Connecticut Avenue, NW, Suite 510
Washington, DC 20036
202-822-9002
iad@thedialogue.org
www.thedialogue.org and www.preal.org

Citation: Ravela, Pedro, Patricia Arregui, Gilbert Valverde, Richard Wolfe, Guillermo Ferrer, Felipe Martínez Rizo, Mariana Aylwin, and Laurence Wolff, *The Educational Assessments That Latin America Needs* (Washington, DC: PREAL, 2009).

ISBN: 978-0-9800777-5-9
First edition
Published in the USA

Book design: Nita Congress
Cover design: Studio Grafik

Contents

1. Introduction

This working paper addresses the need for and uses of large-scale standardized assessments, typically in primary and secondary school, of learning and/or educational achievement in Latin America and the Caribbean. It is aimed at policymakers in the field of education, teachers, academics, business people, members of trade unions and social organizations, and those working in financial agencies and the media. It seeks to contribute to the debate on standardized testing in education systems and to decisions made in that regard.

As used here, **large-scale standardized assessment** refers to assessment that produces comparable information on the performance of students from different cultural and regional contexts, including from different countries. It yields information that provides an overview of the situation in a country, state, or province, even where the sample is not particularly large (for example, 5,000 students).[1]

This document focuses on assessments of **learning**, defined as the change in each student's knowledge and capacities in the course of the school year, and/or of **educational achievement**, meaning the accumulation of knowledge and capacities throughout a student's whole life. The paper does not address other important aspects of assessing educational endeavors, such as evaluations of teacher performance, education policies, and schools or the assessments that teachers carry out in the classroom. It also does not examine tertiary-level assessments or entrance exams for higher education.

This working paper can help those who make education policy decisions understand and analyze various options regarding the purposes and uses of assessment systems and the implications of each, and how to devise an assessment strategy.

All too often, assessment is undertaken in a simplistic and naïve manner, leading to poorly conceived and inadequately implemented assessment exercises with unintended consequences for the education system as a whole—including wasting resources and discrediting assessment in the eyes of teachers. A wide range of issues must be taken into account in implementing an assessment system or reforming an existing scheme. If the assessment initiative is to have any value, the education system in which it is implemented must have a clear purpose, an underlying vision of shared responsibility in education, a high-quality technical design consistent with its aims, a strong motivation to support teachers in their work, and the political will to take the steps necessary to resolve the problems and weaknesses the assessment uncovers. Significant resources must be invested in creating a competent technical unit, and a comprehensive long-term plan must be drawn up; this process takes time and cannot be improvised.

2. The Importance of National Assessments of Educational Achievement

Standardized assessments are used with growing frequency to gain greater insight into processes and results in education systems throughout the region and the world, in countries with very different cultures and with governments of varying political affiliations. This is evidenced by countries' increasing involvement in international assessments such as the Programme for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS), and the Progress in International Reading Literacy Study (PIRLS) as well as regional

[1] The term "large-scale" is not synonymous with census-based assessments.

tests such as the Second Regional Comparative and Explanatory Study (SERCE) in Latin America and the Southern Africa Consortium for Monitoring Educational Quality (SACMEQ) in Africa. It is also apparent in the development of various kinds of national and subnational assessment systems.

In some cases, assessment is spurred by a desire to engage in civic education and consolidate a democratic society. In other cases, the objective is to increase the productivity of the labor force and the competitiveness of the national economy, to take advantage of the opportunities for the overall development of individuals and their participation in the knowledge society, or to use education as a path to securing greater equity and overcoming poverty. Many assessment systems are based on a combination of these concerns.

In almost all cases, it is assumed that assessment can serve

- as a basis for better-grounded education policies,
- as a means of improving the management of education systems,
- as an instrument to foster collaboration and continuous learning within those systems.

The following describes the main contributions of such assessments.

Standardized assessments provide an overview of results for the entire student body. Education is an opaque activity, in the sense that its outcomes are not directly or immediately apparent. Good teachers can determine whether their students are learning and how, but not all teachers have the same assessment criteria. Those criteria are closely related to teachers' professional experience: their general and specific training, their knowledge of the subject they are teaching, their capacity to discern students' processes and difficulties, their familiarity with different kinds of students, and so on. Given the diversity of a country's teachers, an overview cannot be acquired simply by aggregating individual viewpoints, however. Standardized assessment seeks to provide that overview.

Assessments offer information on the real extent of knowledge and the capacities that students achieve, measuring more than just years of schooling. In past decades, the relationship between staying in the education system and developing knowledge and capacities was taken for granted; thus, the indicators used to appraise education systems were related to access (enrollment, coverage, dropout rates, etc.). In that period, the poorest majorities with the least cultural capital either failed to enter the education system or had just a few years of basic schooling. As access to the system has gradually become universal, students from the poorest social sectors who do not speak or write their country's official language fluently arrive at school at a distinct disadvantage and seldom perform at the grade level in which they are enrolled. At the same time, increased access to the teaching profession has not been matched by guarantees of quality in teacher training, a circumstance that casts doubt as to whether such an equivalence exists.

In sum, an increase in the number of years that pupils spend in the system does not necessarily mean that all children and youths are developing the increasingly sophisticated and complex knowledge, attitudes, and capacities needed for their personal and social lives. Assessments seek to shed light on what is happening in this regard.

Standardized assessments help reveal a set of key educational issues. Among other things, standardized assessments provide information on

- the extent to which students are learning what is expected of them when they complete certain grades or levels of schooling;

- the degree of equity or inequity in the acquisition of such learning;
- how achievement levels and equity in the access of different social groups to knowledge evolve over time;
- how and to what extent socioeconomic and cultural inequalities affect students' learning opportunities;
- the range of educational practices currently used in schools and among teachers, and how those practices are related to students' learning in different social contexts;
- how student progress is influenced by teaching conditions (teachers' wages and general working conditions, access to teaching resources, availability of time to prepare lessons and reflect on teaching practices, etc.);
- the effect that investments in educational programs, changes in the education system's structure, curricular reforms, training schemes, the acquisition of teaching materials, and other factors have on educational attainment.

A system for assessing learning and/or educational achievement can provide important information to various social actors. To the extent that the system produces and appropriately delivers information on the above matters, it can be a crucial instrument for improvement, enriching understanding of educational circumstances and decision making in various spheres.

- **Ministry of education officials and other policymakers** can gain a better understanding of problems in teaching and learning; they can be made aware of deficiencies in the context within which teachers discharge their duties; and they can develop relevant policies to support the work of the schools. Assessments also give them a solid basis of empirical evidence on which to evaluate the impact of policies and programs they have implemented, and to gauge the probable effects of those they plan to implement.
- **Principals and teachers** can, on the basis of an external review of educational achievements throughout the system, gain a better understanding of the level at which their students are performing, how they are learning, and the difficulties they are facing. They can learn from the experiences of other teachers and schools working with students similar to and different from their own. And they can make more informed decisions about which aspects of the curriculum to emphasize and can enhance their own methods of assessing student learning.
- **Supervisors and those responsible for training teachers** can use information on systemic educational achievements and difficulties as the basis for making a detailed study of weaknesses in overall educational approaches or specific teaching practices that underlie the learning deficiencies the assessments reveal. They can thereby improve their work in guiding and training teachers. Supervisors can particularly benefit from using assessment data to map schools by student sociocultural composition and educational achievement.
- **Parents**, given appropriate information, can better understand what their children are expected to learn, what they are achieving, and what they can do to collaborate with the school in their children's learning.
- **Citizens** in general may pay more attention to educational issues and problems when they are better informed about what is happening within the education system. They will be better positioned to demand of public officials and teachers that the

To Avoid Confusion…

When national assessment systems are being developed, their role and function must be stated explicitly, along with their limitations. This explicit definition will obviate the risk that they, as well as the information they gather and disseminate, might be misinterpreted.

- **A standardized assessment provides fundamental and vital information on educational quality, but is not a complete indicator of said quality.** Not all the valuable goals of education are included in national standardized assessments. Much important knowledge and learning, as well as many significant attitudes and values, cannot be part of an assessment—because they are difficult to measure—or should not be part of one—because they are relevant to their local context and thus not applicable to all students in a country.
- **Standardized assessment of learning and educational achievement is an essential component of a comprehensive educational assessment system, but it is not the only form of relevant evaluation.** Also important are the assessments that teachers carry out in the classroom and assessment of teacher performance, of educational institutions, of education policies, of resource use, of the relevance of the curriculum, and so on.
- **Assessment is a necessary but insufficient condition for improving education.** There is some evidence that the mere existence and dissemination of information has some effect on certain actors. But assessment is only one of several key elements of education policy; others include preservice and in-service teacher training, teacher working conditions, school management and supervision, curricular design, textbooks and educational materials, investment of resources proportional to the needs of different populations, and concerted action by those responsible for education to resolve any problems uncovered. Significant efforts should be made to ensure that these elements are in proper alignment.
- **Standardized assessment will only have positive effects on education if it is conceived, perceived, and used as a mechanism to ensure the public accountability of all actors involved in education.** There is always a risk that education policy might concentrate on conducting assessments, but that no concrete action is subsequently taken to tackle and resolve the problems the assessments bring to light. Often, authorities confine themselves simply to providing information on the assessment results and transfer all responsibility for solving the problems to the schools and families. At other times, all responsibility is assigned to the teachers, who themselves tend to shift responsibility to the parents or the social context. This "blame game" is unproductive; instead, a vision of shared responsibility for education should be forged.

quality of education given to children and youths be constantly improved, and that the resources devoted to education be used responsibly.

Development of a national standardized assessment system necessitates informed debate as to which aspects of the formal curriculum should be mandatory and to define clearly what all students should have learned at the end of each educational cycle. In Latin America and the Caribbean, most curricula consist of long lists of goals and themes, all of which are desirable but not all of which are feasible. Designing national assessments makes it necessary to determine what should be considered fundamental—i.e., what all students should know and be able to do—regardless of whether this knowledge set is defined as standards, basic skills, achievement indicators, performance levels, learning targets, proficiency criteria, etc.

3. Standardized Assessment in Latin America: Current Situation

National standardized assessment systems experienced robust development in Latin America during the 1990s. In some countries, the systems have been operating continuously for decades, despite shifts in approach or changes in institutional structure. Other countries have experienced a significant lack of continuity and have repeatedly had to start over almost from square one, or will have to do so in the near future.

- Sixteen countries are taking part in the UNESCO Regional Bureau of Education for Latin America and the Caribbean (OREALC/UNESCO) Second Regional Study for third- and sixth-grade primary school students: Argentina, Brazil, Chile, Colombia, Costa Rica, Cuba, the Dominican Republic, Ecuador, El Salvador, Guatemala, Mexico, Nicaragua, Panama, Paraguay, Peru, and Uruguay. The related network of national assessment systems, in which most Latin American countries participate, and which constitutes OREALC/UNESCO´s Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE), has been active in the region since 1997. It meets twice a year and is a useful forum for training and the exchange of experiences in the assessment field.
- Six countries—Argentina, Brazil, Chile, Colombia, Mexico, and Uruguay—took part in PISA 2006, and another three—the Dominican Republic, Panama, and Peru—will be involved in PISA 2009. A PISA Ibero-American Group was recently established, comprising Argentina, Brazil, Chile, Colombia, Mexico, Portugal, Spain, and Uruguay.
- Some countries have taken part (or are now participating) in the studies for mathematics and science (TIMSS), reading (PIRLS), and civic education led by the International Association for the Evaluation of Educational Achievement.

Large-scale assessments in the region are becoming better and more frequent. Changes and improvements in recent years include the following.

- **Greater transparency in the dissemination of results.** Authorities no longer interfere in the disclosure of "unfavorable" assessment results, as those in several countries had done in the past.
- **A shift from normative tests, whose main aim is to establish a comparative hierarchy among students, to criterion-referenced tests, which focus on what students know and are able to do.** Increasingly, too, criterion-referenced tests include a definition of the results that all students should achieve if their performance is to be judged satisfactory.
- **Improved technical/methodological capacities to devise tests and process data.** Efforts have been made to develop tests that assess a broader range of knowledge and capacities, and that include open-ended questions. The use of more sophisticated methods of data processing, such as item response theory and multilevel analysis, has also been growing.
- **Increased attention to the dissemination and use of results.** There is a rising awareness that it is not enough to undertake an assessment and publish a report. A dissemination strategy must also be developed, and a series of reports must be prepared in line with the needs of each target audience. Assessment units are more aware of the appropriate and inappropriate uses of each kind of assessment, although those who make the policy decisions are often not clear about this.
- **Greater concern for research into the factors that affect learning.** Though much remains to be done in methodological and interpretive terms to produce good research studies, there is, in most

countries, an appreciation of the need to conduct more research and to devise hypotheses about how children's opportunities to learn are influenced by the dynamics of education system management, school processes, teaching practices, and education policy decisions.

- **Countries' increasing involvement in international assessments.** This involvement has had positive effects on assessment units and the quality of their work: it has fostered the development and accumulation of technical capacity for assessment, facilitated exchange and the development of a common language among countries, and helped improve the quality of various technical processes (the design of tests, inclusion of open-answer questions, sampling, quality control of the test administration process, data analysis, dissemination methods, etc.).

The primary weaknesses in the region's standardized assessment systems are in the areas of dissemination strategy, use of results, and technical quality of assessments. Despite the aforementioned improvements, a significant number of weaknesses remain to be addressed.

- Often, political authorities demand that large assessments be conducted without clearly defined purposes, with time frames and resources that are insufficient if the assessments are to be technically sound.
- Authorities frequently fail to understand that not just any assessment will serve any purpose, and that an assessment system must be designed carefully for the long term in light of clearly established purposes and uses. Without this definition, good decisions cannot be made regarding the use of censuses or samples, grades and subjects to be tested, periodicity of assessments, types of tests and reporting scales to be used, among other things.
- There should be more public discussion of what should be assessed and which aspects of the curriculum should have been taught to all students by the end of certain levels of education. Absent such a discussion, learning targets and/or standards are unclear, and there can be no clarity for teaching or assessment. This undertaking calls for a broad social debate and for the coordination of technical work between curriculum and assessment units.
- In order to raise the quality of research on factors associated with student learning, it is important to improve the instruments that are applied with the test to collect complementary information, as well as to develop more ambitious research designs, such as longitudinal studies, "value-added" models, qualitative studies, etc.
- In general, the region's education ministries have had only limited capacity to conceive, formulate, and implement policies that respond to the problems uncovered by the assessments. There should thus be stronger links between the assessment units and other important actors in the education field, both within and outside the education ministries. Closer ties among assessment, curricular development, preservice teacher training, and teachers' professional development are also needed.
- The activities undertaken for the dissemination and use of the assessment results continue to be inadequate, especially in terms of ensuring that teachers understand and use the results and that the results are incorporated into the school culture.
- Most countries persist in disseminating results by type of school and/or by individual schools, states, or provinces without due consideration for the sociocultural contexts in which the institutions and educational subsystems operate. This practice gives rise to flawed interpretations and conclusions regarding the educational effectiveness of those in-

The Risk of Poor Assessments

Only an assessment that is technically sound and whose results are disclosed and used appropriately can have an impact on improving learning. An assessment whose results are little known and little used is obviously a futile exercise and a waste of resources.

More serious are those circumstances in which the results of technically deficient assessments are broadly disseminated, or when assessments are used for purposes other than those for which they were conceived and designed. These scenarios can be harmful to the education system itself. Moreover, to simply carry out an assessment and disseminate the results as an accountability mechanism, irrespective of the quality of the assessment, can cause assessment systems to suffer setbacks and make it impossible to hold a serious discussion about the advantages and disadvantages of accountability for results.

stitutions and subsystems. A proper appraisal calls for value-added assessments, involving two evaluations of the same population at different times. The region has only minimal experience in this area.

- Many countries continue to exhibit significant technical weaknesses in their assessment design: most of the questions in their assessments are overly simplistic, with few if any questions devised that allow complex cognitive capacities to be measured. The assessments focus on the educational achievements of a specific grade, which makes it difficult to determine what has been learned by those who still fail to attain what is expected of that grade. There are deficiencies in devising samples and in estimating and reporting measurement errors. Weaknesses also persist in conducting assessments that are comparable over time.
- There is a shortage of professionals who are qualified to design and carry out these kinds of assessments, a circumstance exacerbated by turnover among the countries' technical specialists, often for political reasons. This lack of expertise and continuity hampers the accumulation of knowledge and experience in the region. Many countries have had to start from scratch in establishing their assessment systems, years after having had a system in operation and having dismantled it.

4. Purposes and Uses of Assessments

Whether an assessment system is being implemented or reformed, countries must evaluate certain basic options in light of what the system is expected to deliver.

The first step is to define the purpose of the system and the use to be made of its results. Several options should be considered; they are not necessarily mutually exclusive. A system can in fact combine several of them, but it is important to be aware that each requires a particular design and has different technical demands and costs. The most basic decision is to determine whether the national assessment system is to be diagnostic in nature or whether it is to certify achievement.

- **A certification assessment or graduation examination** is a means of determining students' educational achievements. Here, the main goal is to determine which students have acquired the requisite knowledge and reached the requisite performance level to complete a course or grade and thus to pass or fail.
- **A diagnostic or formative assessment** can focus on students, schools, or the education system as a whole, and entails no direct consequences for

students. Its main goal is to provide quality information to enrich the perceptions, decisions, and actions of various actors (authorities and technical staff, supervisors, principals, teachers, students and their families), with a view to improving teaching and learning.

A proper balance must be struck between the implementation of national tests and participation in international tests. National tests can offer a better view of what students have learned as compared with what they have been taught, while international tests can reveal what students know and can do compared to their counterparts in other societies, thereby enriching the debate on the national curriculum and approach to teaching. Countries should carefully evaluate which international tests they might choose to participate in, mindful of what each test seeks to assess and its relevance to national objectives. It seems advisable to participate periodically in at least one regional or international test, in light of each country's priorities.

4.1 Assessments to Certify What Students Have Learned

An assessment system to certify what students have learned by means of a high-quality, national examination has advantages in terms of transparency and accountability for results. It is common for two students who have completed the same level of education in two different regions of a country to have very different levels of knowledge. Assessment systems designed for certification purposes clarify the value of educational qualifications in the eyes of society. Moreover, such systems make both teachers and students themselves responsible for acquiring the knowledge that the test assesses, which has positive effects on learning. Such graduation exams are much more appropriate for the higher levels of the education system than the lower, especially at the end of secondary school.

An assessment system for certification purposes has some costly requirements. First, the tests must be census-based and each student must have more than one chance to take them, thus requiring the conduct of several assessments every year. Additionally, because the tests must have wide curricular coverage, they must be extensive, covering several subjects or disciplines. They should also, to the extent possible, include constructed response questions, which entail significant codification costs.

An assessment system geared to certification can trigger considerable tensions which must be anticipated. If the graduation tests are demanding, there might be a very high failure rate. This would mainly affect the more vulnerable social sectors and could intensify such problems as dropout rates and youth unemployment, making such tests unsustainable from a social and political standpoint. Consequently, this type of assessment—like all assessments—must be accompanied by complementary and remedial teaching activities. If high standards are proposed, systemic responsibility must be taken to provide students with every opportunity to learn, including the provision of textbooks, materials, facilities, and teacher preparation. The following measures will also be useful in implementing these assessments:

- Establish that standardized assessment is only one part of the process of certifying student learning (for example, 40 percent of the final grade). The rest of the score could be derived from teacher assessments of students. Even though teachers' assessment criteria are diverse, this approach combines external and internal assessment, and gives teachers an "outsider's" perspective regarding what is happening in their schools, thereby allowing them to reflect on their own evaluation criteria.
- Establish a period of transition to universally applicable standards. During that period, the main

focus would be on the extent to which the students in each school had improved or made progress relative to their previous performance and not solely on achievement of the standard for certification.

The worst way to approach the issue would be to set a fixed failure rate—for example, to mandate that no more than 10 percent of students should fail. Such an approach entails the use of simpler tests and thus sends the wrong signal to students, families, and educators about academic expectations.

4.2 Diagnostic and Formative Assessments

A formative assessment system with no direct consequences for students has advantages in terms of cost, possibilities for designing and setting high standards, and contributing to an assessment culture. The costs of applying such tests can be lower, since they can be taken by a sample of students in key grades and at certain multiyear intervals. The tests can be matrix-type, in which not all students answer the same questions but rather blocks of questions. This method allows a very large number of questions to be used and facilitates a more detailed analysis of the various aspects of the curriculum. Such systems make it possible to set demanding standards or expectation levels without causing widespread failure. Moreover, they help promote an assessment culture and allow technical capacity to be constructed, so that when consideration is given to the possibility of establishing a system with consequences, the conditions are in place to do it properly. One of the alternatives to be considered in using formative tests is to release an entire test so that it can be applied autonomously by the teachers in order to help them identify the difficulties of individual pupils and enrich their repertoire of assessment instruments.

The main problem with diagnostic assessments is that they may have no impact if complementary actions are not taken. Diagnostic and formative assessments can be of little use if they are not matched by a precise strategy and significant investment to ensure that results are disseminated and used in later educational activities, since their effectiveness depends on the various stakeholders receiving, understanding, and using the results. In this regard, the following considerations should be kept in mind:

- **If the results of such assessments are to have an impact on education policies**, time must be spent in analyzing and discussing the results in various offices of the education ministry and among other relevant stakeholders, in understanding the problems and weaknesses that the results bring to light, and in devising appropriate activities and investments to tackle those problems. The authorities must be willing to submit their policies and decisions to public scrutiny; hence the need to invest in appropriate and ongoing communication of results to the public.
- **If the results are to have an impact on teaching practices**, time must be spent in analyzing and interpreting their didactic implications: if students are unable to resolve certain situations, what is being done deficiently or incorrectly in the classroom, and what should be done differently? This kind of analysis should be undertaken by individuals who specialize in teaching the subjects under assessment and by teachers, thereby creating permanent forums for in-service training and joint work within the schools. Teachers must be able to analyze the greatest possible number of items in order to determine which reveal a significant obstacle to the development of new concepts or capacities. However, it will always be necessary to keep some assessment items confidential so as to enable comparable assessments over time.
- **If the results are to have an impact on students' motivation and families' attitudes toward learn-**

ing in school, students and families must be given appropriate and comprehensible information as to what is regarded as crucial for students to have learned in each grade or educational level and on the actions that might be conducive to attaining such learning.

Diagnostic tests can be sample- or census-based, depending on the strategy for educational change. Whether the tests are sample-based or census-based has various implications; it is also possible to combine a test for controlled samples with a census-based distribution of tests that can be applied autonomously by the schools. The purposes of the latter are formative, facilitating analysis of results and identification of students who need additional support.

- **Sample-based tests** provide an overall diagnosis of the system. Care must be taken in devising the sample, so as to secure representative information for the levels of disaggregation at which actions and decisions are to be taken (regional, provincial, and municipal; urban and rural; indigenous schools; etc.). The impact of these test results depends mainly on the education policy measures taken at the central level and on an appropriate outreach strategy targeting all schools.
- **Census-based tests** provide information on each school and even on each pupil. The impact of their results depends on conveying the information to each educational community with a focus and format that helps encourage greater participation and commitment at the local level. The information might also be highly useful in focusing policies on districts or schools with more serious problems, since census-based tests offer a "map" of the results of every school, area, province, type of school, and so forth.

Diagnostic Assessments Should Take a Broad Perspective

Tests should not be confined to assessing the knowledge and skills corresponding to a single grade (that for which the test is designed), but rather should take a broader perspective of performance levels—from the most basic to the most complex—across several grades. This approach helps identify what students have learned in previous grades and what they need now. Teachers can thereby note and rectify the weaknesses of learning in previous courses that hinder students from advancing. The results will be useful not only for teachers working in the grade under assessment, but also for those teaching earlier grades. The emphasis should not be on informing the teachers that students have "passed" or "failed," but rather on conveying that they are at different points on a continuum of learning, along which everyone can and must advance.

4.3 Assessments and Incentives

Some standardized assessment policies are geared to establishing economic incentives in light of assessment results or to fostering a competitive market among schools. These policies are of three types:

- Using the results to draw up rankings of schools and making them public as a means of encouraging schools to take responsibility for their results, helping families make informed decisions about which school they want their children to attend, and promoting interschool competition to achieve the best results.
- Using the results to give economic incentives to the schools that achieve the best results or improve on their results in previous assessments.
- Using the results as an indicator of the quality of each teacher's work and as a criterion for offering economic incentives.

These approaches sometimes result—deliberately or not—in the state relinquishing its responsibility for the results of the education system. Sometimes, by confining itself to carrying out assessments, delivering the results, and creating incentives on the basis of the results, the state transfers responsibility for the results to schools and families, as if it were a matter between private actors. The state does this rather than endeavoring to create the conditions wherein teaching is made effective by providing the necessary resources, putting in place a properly trained teaching staff, and establishing mechanisms for evaluating and guiding the work of the schools. This approach does not take into account the complexity of the educational endeavor, especially in socially disadvantaged environments, and it disregards the need to invest in capacity building as a key tool in improving teaching and learning.

4.4 Disseminating and Using Results

Responsibility for results should be shared among the various stakeholders, including national, regional, and local authorities; the teaching staff; and students and their families. Care must be taken to avoid using the results for the deliberate or implicit purpose of assigning sole blame or responsibility to certain actors.

Responsibility for results requires a proper balance between the demands made on schools and teachers and the support they are given. Schools and teachers should take responsibility for ensuring that all students learn what is expected of them. At the same time, the authorities have the duty of establishing the support policies needed to allow schools and teachers to do their work properly. Making heavy demands of schools and teachers without providing the corresponding support can only cause ill-will and discouragement. On the other hand, providing support without making the corresponding demands can cause complacency.

It is not appropriate to use the results of standardized assessments as the main indicator of the quality of the work of the teacher or school. This premise is particularly important where the assessments do not control for other factors within and outside the education system, and where it is not kept in mind that learning also depends on student motivation and personal effort. Because the results of standardized assessments are not the sole indicator of the quality of education, they must be considered in conjunction with other pertinent matters such as educational attainment, the relevance of what is being taught, the development of values and habits, and civic education. Efforts must be made to avoid identifying the term "quality" with the results of standardized tests.

To facilitate understanding and the use of results, it is not enough to offer only numerical data. The various stakeholders must understand the kinds of tasks students should be able to undertake in the tests. Concurrently, however, the question set must be kept confidential so it can be used in future assessments to ensure comparable evaluations over time. While it is not intended that the teachers use standardized tests to assess their students, knowledge and understanding of these assessments can help improve their own evaluation methods and develop an assessment culture.

When the differences in results are reported in terms of the sociocultural composition of the student body, care must be taken to avoid creating a system of differentiated achievement expectations for diverse social groups. Education policies should make distinctions so as to create the right conditions in which to teach the least advantaged groups. No attempt should be made to use the results in such a

way as to encourage schools—directly or indirectly—to select students with a view to improving their results.

4.5 Some Cautions about Comparing Results across Schools

Although the results of standardized tests do not provide an exhaustive picture of a school's educational quality, they do provide important information on the performance levels achieved. The information on two standardized test results—normally, language and mathematics—is not of itself an evaluation of the schools' "educational quality" and should not be presented to the public as such. The "quality" of a school includes other relevant matters that are of importance to teachers, students, and families—such as emotional development, interpersonal relations, civic education, and the inculcation of values. Nonetheless, comparative data on the performance levels achieved by students in a range of schools can be useful to the teaching staff, inasmuch as the information enriches their perception of their own work, allowing them to locate their students' achievement in the context of that of students in other schools.

If a comparison of student performance levels is to be valid, the students' social background must be taken into account. Schools' academic results must be compared with those of schools with a similar social composition. This is because the challenges and difficulties of teaching students from disadvantaged backgrounds (or those whose native tongue is an indigenous language) are very different from those involved in teaching students from family backgrounds marked by complete secondary and/or tertiary education. School dropout rates should also be kept in mind, as should student selection policies, since a school can improve its results by excluding students who have difficulties.

Between-school comparisons should take account of the difference between measuring "educational achievement" and "learning." Strictly speaking, if the aim is to provide schools, families, and/or the authorities with information on the teaching capacity of teachers and schools, assessments must measure both progress made by students throughout an academic period (learning) and the final result (achievement). The differences between the two terms are as follows:

- **Learning** can be defined as the change in each student's knowledge and capacities throughout the school year. Measuring it requires two tests, one at the start and another at the end. This approach makes it possible to determine the progress that each student has made.
- **Educational achievement**, by contrast, is measured using a single test and reflects the accumulation of knowledge and capacities throughout a student's whole life, including the family's cultural capital and the student's experiences in other schools or with other teachers.

Students' learning and educational achievement depend not only on what teachers and schools do, but also on the effort made by the students themselves, the support that the families give to their children's schooling, the community and cultural context, and education policies. Problems in education cannot be resolved by appealing solely or mainly to market mechanisms.

Results expressed as school rankings should be viewed with caution. Most school rankings give a false impression of standing. One school might be in 1st place and another in 40th, but the difference between their averages might not be statistically significant. Thus, it cannot be said that one average is really higher than the other, because the differences

fall within the assessments' margin of error. But even when the difference in the averages is statistically significant, it might be irrelevant in terms of the percentage of students who attain the expected performance levels.

4.6 Checklist for Decision Making

Given the various options for devising and implementing an assessment policy, it is essential to determine the assessment system's characteristics before putting it into effect (or modifying an existing one). The following checklist can be helpful in assessment system decision making.

- What is the purpose of the assessment? Who will use the results and to what ends? What as yet unknown information will the assessment provide?
- What are the units of analysis for the results report: individual students, class groups/teachers, schools, types of schools, subnational governing bodies, the education system?
- Based on the defined purposes of the assessment, which is more appropriate to assess at the end of given grades or cycles, learning or educational achievement?
- What consequences will the results have, and for whom?
- According to the defined purposes, is a census-based approach necessary, or are sample-based assessments sufficient?
- What grades and disciplines should be assessed?
- How often is it necessary and appropriate to undertake the assessments?

The answers to these questions should be integrated in a clear and explicit assessment plan for the short, medium, and long term. In drawing up the plan, the

"Commandments" for Making Policy Decisions about the Assessment System

- **"Do no harm."** One of the first precepts of the Hippocratic Oath is that physicians will refrain from doing anything that could harm their patients. Similarly, when planning an assessment system, it is important to consider the risks of unintended and damaging effects that the selected assessment strategy might have on the education system whose improvement is sought.
- **"Everything in moderation; nothing in excess."** Assessment cannot take primacy over education. Too many assessments can harm the health of the education system, especially if countries only conduct assessments and fail to devise policies that respond to the problems thereby uncovered.
- **"Do not assess in vain."** Collecting information every year and never analyzing or using it should be avoided. If an assessment system is to have an impact, the assessments should be carried out at intervals that guarantee that the data can be analyzed, discussed, understood, and used. It takes time to absorb new information and translate it into decisions and actions. Changes in the education system need even more time.
- **"Dress me slowly, for I am in a hurry."** Decision makers in education ministries must reject the false belief in easy, quick solutions. A serious assessment program cannot be established in three months. There is no circumstance or window of opportunity that justifies it: sooner or later, the consequences of improvisation will become apparent. Assessment demands careful reflection on its purposes and uses, public discussion of what needs to be assessed, the involvement of various actors and dialogue among them, the creation of technical teams that are competent in various matters, and prior information that motivates the actors to become involved in the assessment.

monetary costs of each option and the human resources needed to implement it properly should be taken into account. Most particularly, a balance must be struck between the investment required to collect the information and that needed to disseminate and use the results. Many education ministries devote significant sums of money each year to gathering large amounts of data that thereafter are scarcely analyzed, disseminated, or used. It is pointless to undertake an initial assessment exercise if no long-term work plan has been prepared.

5. Technical Quality Challenges of Assessments

Once the assessment policy has been defined, it must be implemented in line with various appropriate standards of technical quality. This entails addressing the following challenges.

Devise a referent or conceptual framework that stipulates precisely the knowledge and performance deemed appropriate at the end of the grade or educational cycle to be assessed. While such an endeavor calls in the first place for a political debate and political decision making, the definitions thereby decided upon must then be adequately translated into technical specifications and standards. The education policy debate must be fueled by information on recent conceptual developments regarding teaching and student performance in the disciplines under consideration.

Include activities with various levels of complexity in the tests. Thought should be given to including activities that call for sufficiently complex cognitive capacities that are appropriate to the challenges of the knowledge society, as well as simple activities that reveal the level to which less advanced students have progressed. The activities must be guaranteed to possess a series of psychometric properties, and controlling them demands pilot tests and careful analysis. To the extent possible, it is also important to expand the use of constructed response questions, with the twofold aim of assessing more complex capacities and improving the links between external assessment on the one hand and teaching and school culture on the other.

Design the tests by properly integrating activities in blocks and test booklets. This is a significant and complicated technical challenge, calling for specialized knowledge and experience. Particular attention should be paid to the decision of whether to use classical theory or item response theory in devising and analyzing the tests; this in turn calls for modern processing programs, well-trained analysts, rigorous analysis, and high-level advice.

Define the cut-off scores that set the limits between performance levels in a test. Authorities must define a methodology to establish which of the performance levels should be deemed acceptable for a student completing the grade or educational cycle under assessment. An acceptable performance cannot be defined automatically as equal to 51 percent or more of the maximum possible score on a test.

Devise samples that are suitable for the purposes of the assessment. The aim here is to avoid assessment exercises that are larger and more costly than is strictly necessary and that, at the same time, have an appropriate degree of precision. In this latter regard, the margins of error in the assessment should be estimated and reported.

Determine how to equate the assessments. Equating the assessments refers to the methodology used to establish comparability among the results of tests applied in different years. This is one of the most important technical challenges assessment systems face

if they seek to provide information on progress and setbacks in educational achievement over time. The process is crucial if the assessment is to be able to indicate that any possible variations respond to changes in real educational circumstances and not simply to changes in the measuring instruments. Answering this challenge means taking into account statistical considerations, assessing the same knowledge and skills in each subsequent assessment, and keeping the structure and length of the test comparable over time.

Produce longitudinal data that reveal how the learning of the same group of students has evolved over time. Production of such data demands more than measuring achievement over time. It assesses "learning" as change and is more apt for establishing which school processes influence learning. This approach provides more relevant information for education policymaking and for research.

Establish mechanisms to monitor compliance with the standardized conditions that should prevail during test application. Compliance with the conditions in which the tests should be applied involves complex technical matters that are often neglected. These issues are crucial to ensuring that the information obtained is reliable and comparable and relates to such issues as the following:

- The quality of the test administrators' training
- Establishment of quality controls during test administration
- Student motivations to take the tests
- Distribution logistics
- The safe return of test materials

Combining assessments with qualitative studies. It is important to combine standardized national assessments, which offer an overview of the whole system, with qualitative studies that reveal more about processes in the schools and classrooms. This combination of approaches is the best means of providing rich and complex information to guide thinking about education policies and teaching practices.

Ensuring Needed Transparency

Technical processes need, above all, transparent and accessible information. Particular emphasis should be placed on documenting the technical procedures followed in

- designing instruments,
- estimating the precision of measurements (and, consequently, their margin of error),
- designing samples and reporting actual response rates,
- applying the tests and controlling the quality of their administration,
- defining performance levels and cut-off scores,
- equating the results and making them comparable with previous assessments.

6. Establishing Assessment Units

Countries need assessment units with the requisite capacities and resources in order to move forward with an appropriate policy for assessing educational achievement or learning.

It takes at least two to three years to establish a serious assessment system. That much time is needed to move ahead with the following basic processes:

- To discuss, define, and make publicly known the purposes of the assessment system, the kind of consequences it will have, its expected uses, and what is to be assessed

- To design a long-term assessment plan
- To assemble technical teams with the necessary range of skills (design of tests and questionnaires, knowledge of the disciplines to be assessed and of their didactic requirements, curriculum and standards, sampling, logistics and quality control of the administration process, gathering and cleaning data, processing and analysis, scale construction, sociocultural contextualization of the results and analysis of related factors, data interpretation in terms of education policies and teaching practices)
- To devise instruments, test them in a pilot phase, and ensure external supervision of these processes

The technical staff of the assessment units must be stable over time. It takes about 10 to 15 years of planning to develop an assessment system. A high rate of turnover among technical staff causes a loss of knowledge and accumulated experience in a complex field and can discredit assessment processes in the eyes of educators and the public.

Assessment units must be independent in reporting the results of their work. As with units that provide social and economic statistics, those responsible for educational assessment and for disseminating test results cannot be dependent on the time frames and interests of political parties. There has been much discussion of whether assessment units should be located within or outside of education ministries. The main argument for the latter approach is precisely this need for independence and transparency. Nonetheless, there have been examples in the region of countries with stable and independent units within education ministries, and other cases where the units have been unstable and unable to consolidate their work even though they were part of an external institution. In fact, the institutional location is not as important as the culture of continuity and transparency created around assessment. Such a culture is achieved when assessment has a clear mandate and a solid structure, which necessitates that the assessment system be underpinned by some kind of legal statute. One of the approaches to be considered is to establish this by law, since it calls for broad agreement (that reaches across party lines, if possible) that allows a long-term educational assessment plan to be put in place. As in the economic arena, there must be a certain level of stability: if assessment policies change constantly, distrust grows and credibility is lost.

A solid institutional structure requires independence and pluralism among government bodies and technical assistance agencies, an appropriate budget, and human resources that guarantee the unit can function to the necessary degree of technical quality. The assessment unit's independence should not cause it to become dissociated from education policy. On the contrary, assessment must respond to a political-educational project with widespread support and should remain closely linked to other key areas of education policy such as teacher training, curricular development, planning and project design, program evaluation, and research.

If standardized tests are to have an impact on education policies and practices, new interfaces and working methods are needed that ensure that the various actors and decision-making spheres are aligned. The assessment units must assume that their work consists of more than producing data. If their mission is to provide information to other actors, they need trained individuals with the time to establish a dialogue with other agencies and actors responsible for the following:

- **Analysis of education policy**, with a goal of improving the design of the assessments, as well as processing plans and reporting results, and taking into account some important matters for education policy (for instance, the sample can be designed

in such a way as to make it possible to assess the impact of specific policies or programs on certain groups of schools)

- **Didactic analysis**, with a goal of interpreting the results and student learning problems from the perspective of teaching and the didactics of the discipline under assessment, preparing reports with didactic significance for teachers, designing in-service training programs on the basis of the results, and reflecting on the links between standardized assessments and classroom evaluations

- **Communication**, with a goal of designing a range of results reports in various formats and styles tailored to and understandable by various audiences

Good assessment requires investment. It is better to do no assessment at all than to do one that is poor or inadequate. Also, it is better to have a modest assessment system whose costs are affordable and sustainable over time rather than an extensive and sophisticated assessment that can only be carried out once and never repeated.

Investing in assessment should be seen in terms of the use to be made of the results rather than on the basis of other indicators, such as cost per pupil. The costs of assessment are low relative to national budgets and other investment possibilities. But any investment in assessment, be it high or low, is worthless if the results are not put to use.

7. Ten Recommendations on the Assessments That the Region Needs

1. **Assessment must be regarded as an element linked to others in a wider set of education policies and actions.** Assessment in and of itself does not produce improvement. There must be stable links between the domains of assessment and those of curriculum development, teacher training, research, policy design, communications and outreach, among others.

2. **Assessment should reflect coherent and comprehensive consideration on the state of education and the means of improving it.** This process should begin with public consultation and debate as to what students should learn and the purposes and consequences of assessment. It is crucial to have a constructive public discussion of the results, with a view to tackling deficiencies and inequities in students' access to knowledge. This calls for investing in communication and outreach as much as (if not more) in the assessment itself—before, during, and test administration.

3. **Assessment should help develop a sense of shared responsibility for education as a public good.** It should foster all stakeholders' commitment to education in line with their position in the system and their area of activity. Efforts must be made to avoid using assessment as a means of assigning blame to specific actors for problems uncovered.

4. **The region's assessment systems should gradually expand the range of educational objectives that are subject to appraisal.** Civic education and other subjects apart from language and mathematics should be included in order to cover a broader spectrum of competencies and capacities than is currently captured.

5. **The region's assessment systems should gradually design evaluations of student progress over time**, since these can provide additional information on the impact of education policies, actions by the schools, and teaching practices on student learning.

6. **An assessment system is a long-term undertaking, and thus requires commitment on the part of the state and careful planning of its design.** Decisions must be made about its purposes and consequences, the curricular areas to be assessed, topics and grades to be covered, and the periodicity of assessment, among other things. Because careful planning takes time, it is not advisable to try to implement assessment systems over a short period.

7. **A good assessment system needs investment**, primarily in terms of establishing qualified teams as well as providing sufficient economic resources to ensure proper implementation of all the processes involved.

8. **The assessment system should be fully transparent** with regard to the results and accountability to society.

9. **Education ministries must assume a serious and consistent commitment to assessment results**, which entails fostering dialogue on the problems uncovered and the means for approaching them, devising appropriate strategies to resolve those problems, and investing the necessary resources.

10. **The assessment system should be evaluated periodically**, with the aim of analyzing the technical quality of the information it provides and its relevance for various educational and social actors.

PREAL

Partnership for Educational Revitalization in the Americas
Programa de Promoción de la Reforma Educativa en América Latina y el Caribe

PREAL was established by the Inter-American Dialogue in Washington, D.C., and the Corporation for Development Research (CINDE) in Santiago, Chile, in 1995 as a multiyear initiative to build a broad and active constituency for education reform in many countries. It has become the leading non-governmental voice on education in Latin America and a strong advocate for involving leaders from civil society in education reform. Most of PREAL's activities are implemented by a region-wide network of expert public policy and research centers working to promote education reform.

PREAL seeks to improve the quality and equity of education by helping public and private sector organizations throughout the hemisphere promote informed debate on education policy, identify and disseminate best practices, and monitor progress toward improvement. PREAL's activities are made possible by the generous support of the American people through the United States Agency for International Development (USAID), by the Inter-American Development Bank (IDB), the GE Foundation, the International Association for the Evaluation of Educational Achievement (IEA), and the World Bank, among others. The contents of this publication are the responsibility of the authors and do not necessarily reflect the views of PREAL or any of its donors.

INTER-AMERICAN DIALOGUE

Inter-American Dialogue
1211 Connecticut Ave., NW, Suite 510
Washington, D.C. 20036 USA
Tel: (202) 822-9002
Fax: (202) 822-9553
E-mail: iad@thedialogue.org
Internet: www.thedialogue.org & www.preal.org

CINDE

Corporación de Investigaciones para el Desarrollo
Santa Magdalena 75, Piso 10, Oficina 1002
Santiago, Chile
Tel: (56-2) 334-4302
Fax: (56-2) 334-4303
E-mail: infopreal@preal.org
Internet: www.preal.org