redie
ISSN: 1607-4041

Please cite the source as:

Martínez Rizo. (2009). Formative classroom assessment and large-scale assessment: Toward a more balanced system. *Revista Electrónica de Investigación Educativa, 11* (2). Retrieved month day, year, from: http://redie.ens.uabc.mx/vol11no2/contents-mtzrizo2.html

Revista Electrónica de Investigación Educativa

Vol. 11, No. 2, 2009

Formative Classroom Assessment and Large-Scale Assessment: Toward a More Balanced System

Evaluación formativa en aula y evaluación a gran escala: hacia un sistema más equilibrado

Felipe Martínez Rizo
fmrizo@prodigy.net.mx
Programa de Doctorado Interinstitucional en Educación
Universidad Autónoma de Aguascalientes

San Cosme 108, 20010
Aguascalientes, Aguascalientes, México

(Received: May 21, 2009; accepted for publishing: July 21, 2009)

Abstract

Given the proliferation of large-scale standardized tests that has occurred in Mexico in recent years, this article constitutes a review of the international literature on the subject for the purpose of reflecting on the possible consequences of this phenomenon and exploring the progress of alternative assessment approaches. It also reviews the development of concepts related to formative classroom assessment, and summarizes current thinking on this subject. It emphasizes the importance of such approaches for improving educational quality. In conclusion, it argues that it is necessary to move toward assessment systems that combine large-scale assessment and classroom assessment in a more balanced fashion.

Key words: Educational assessment, standardized tests, formative assessment.

Resumen

Ante la proliferación de pruebas estandarizadas a gran escala que ha tenido lugar en México en los últimos años, el artículo constituye una revisión de la literatura internacional, para reflexionar sobre las posibles consecuencias de ese fenómeno y explorar los avances de enfoques alternativos de evaluación. Se revisa también el desarrollo de las concepciones relativas a la evaluación en aula con propósitos formativos, y se sintetizan las ideas actuales al respecto. Se subrayan la importancia que tales acercamientos pueden tener, por lo que se refiere a la mejora de la calidad educativa. Para concluir, se sostiene que es necesario avanzar en dirección de sistemas de evaluación que combinen de manera más equilibrada la evaluación a gran escala y la evaluación en aula.

Palabras clave: Evaluación educativa, pruebas estandarizadas, evaluación formativa.

Introduction: Learning assessment and standardized tests

Learning assessment has an ancient history. China began to apply tests to large numbers of people more than one thousand years before Christ (Oakes and Lipton, 2007). Much later, in the sixteenth century, Jesuit schools initiated a tradition that by the nineteenth century had evolved into *essay type* exams such as the German *abitur* and French *baccalaureate* tests.

In elementary schools learning assessment was systematized later, since educational systems at this level were only consolidated after the Industrial Revolution and the Enlightenment, when it was deemed necessary for all future citizens to at least know how to read and write. Before that, children in wealthy households learned their ABCs from private tutors, or in small parochial or guild schools. The number of students was small and there was no concept of grades. Evaluations did not involve the use of systematic procedures; all that was required was the judgment of the teacher, who did not need to use any special instruments; his daily observation of the progress of each of his students was sufficient.

When the children that were learning to read and write were a minority, their skills levels were also less heterogeneous than today and the quality standards used implicitly by the teacher in making evaluative judgments were relatively simple. With more generalized access to education, students also became more heterogeneous, and it was more difficult to maintain comparable quality standards.

Starting in the nineteenth century, the United States developed an educational system of mass coverage, not just at the primary level, but also secondary level as well as higher education. Therefore, it is not surprising that the earliest large-scale assessments emerged there, in 1845, with the application of history exams to more than 500 students in Boston. In 1895, Rice applied spelling tests to 16,000 students, and in 1897, he tested 13,000 students in arithmetic and 8,300 in reading (De Landsheere, 1986/1996).

In 1890, J. McKeen Cattel published his article *Mental tests and measurements*, a seminal text in which he invents the word *test*. Binet developed intelligence tests,

which were then adapted by Terman at Stanford in 1916, and extended with the Army Test in 1917 (De Landsheere, 1996).

With the development of psychometrics, in 1925 the College Board—a specialized agency that was created in 1900 to develop common entrance exams for a group of universities on the East Coast of the United States—was able to develop aptitude tests (as opposed to tests of knowledge), which went beyond the memorization of isolated facts and focused on the evaluation of basic intellectual abilities. From the 1920s on, work in this field was conducted at Princeton University and in 1948 the office that was in charge of the development of tests was separated from the University, to become the Educational Testing Service (De Landsheere, 1986).

During the second half of the 20th century, American College Testing (ACT) and the University of Iowa also developed important tests. Until that time, however, similar progress was almost exclusively limited to the English-speaking world, to the extent that psychometrics came to be considered an American discipline. This situation became so pronounced that, in 1931, hearing the participants in a congress refer to psychometrics as American, E.L. Thorndike protested, saying "it would be more in the interests of science and of our own comfort, if standardized tests were not called 'American examinations'" (Joncich, as quoted in De Landsheere, 1996).

I. Prevalence and extent of large-scale assessment

The pioneers of standardized tests were convinced that schools had serious problems of quality and that teachers' evaluations also manifested considerable deficiencies. In consequence, they sought to develop instruments for comparing the performance level of students from different schools. Thorndike believed that such tests would remedy the *scandalous lack of reliability in the tests used by teachers* (Shepard, 2006, p. 623).

The advantage of the comparability of results offered by the new tests was appealing, but their limitations were apparent from the beginning. In 1923, B. D. Wood complained that standardized tests measured only *isolated facts and pieces of information, rather than reasoning ability, organizational skills*, etc. From the earliest years of standardized testing, Ralph Tyler also stressed the importance of seeing such testing not as a process separate from teaching, but as an integral part of it (Shepard, 2006).

The content of the texts on assessment used in institutions for training teachers shows that the prevailing idea was that the tests that teachers used in the classroom should be replicas of the large-scale tests. Therefore, teachers should learn to devise structured questions and to analyze the results of instruments developed statistically from such questions. Moreover, they should pay attention to the validity and reliability of these tests, just as in the case of large-scale assessments (Shepard, 2005).

Several events contributed to creating a climate of concern about the quality of education offered to children by American schools at the beginning of the second half of the 20th century. These events included the impact of the launching of Sputnik in 1957 by the then Union of Soviet Socialist Republics (URSS); the Coleman report in 1966; and the downward trend of average scores obtained on the Scholastic Aptitude Test (SAT) each year by aspiring college students.

The National Defense Education Act, in 1958, shows the place of education in the reading of the Sputnik launch, in the context of the cold war (Mathison and Ross, 2008). In subsequent years, the legislatures of California, Florida and Oregon determined that students should be assessed through tests based on minimum performance standards (*minimum competency testing*), as an important part of their strategies for improvement.

By 1982, 42 of the 50 American states had mandatory programs of this type. With more generalized implementation, *minimum competency testing* was often carried out improperly, with the result that its impact was reduced and expectations for the testing were not met (Baker and Choppin, 1990).

Publication of the report *A Nation at Risk*, in 1983, showed America's continued concern for educational quality, from a national security perspective. With it the *educational standards* movement began, going on to gain strength during the 1990s (Mathison and Ross, 2008).

In 1989, at the so-called Education Summit in Charlottesville, the governors of the 50 U.S. states adopted a set of goals to be reached by the year 2000. The third goal declared that by that date "American students will leave grades 4, 8 and 12 having demonstrated competency in challenging subject matter including English, mathematics, science, history and geography" (Mathison, 2008, pp. 8-9). In 1990, with the support of federal funding, procedures for reaching these goals were established, and the National Education Goals Panel and the National Council on Education Standards and Testing were established.

Concern about the quality of education was not confined to the United States. The events that deepened that concern, in particular the launching of Sputnik, also produced reactions that led to the emergence of international assessments throughout the second half of the twentieth century. Even when each country had a national system of evaluation, the comparison of results didn't necessarily follow, given the differences in structure, curriculum and school calendars of educational systems, in addition to differences in content, degree of difficulty and the approach inherent in the testing instruments themselves. For this reason, the groundbreaking work of the International Association for the Evaluation of Educational Achievement (IEA) was remarkable (Postlethwaite, 1985; De Landsheere, 1994).

II. Proliferation of large-scale assessments in the twenty-first century

One consequence that has ensued from the decentralization that has characterized the American educational system is that the large-scale assessments used in each state cannot be compared. The National Assessment of Educational Progress (NAEP), established in the late sixties (Walberg, 1990), provided reliable results in certain subjects and grades at the national level, but not at the individual, school, district or even state level. To obtain reliable results at the school level, other solutions were sought, such as the Voluntary National Test, proposed by President Clinton, or the web-based computerized adaptive testing system proposed by the Rand Corporation (Klein and Hamilton, 1999).

At the beginning of 2002, President Bush promoted new education legislation at the federal level, known as the *No Child Left Behind Act* (NCLB). This legislation meant significant changes in educational policies in general, and particularly in regard to the assessment of student performance.

Through a series of different measures, the act aims to modify in a fairly short period (twelve years, that is, by 2014) the state of American education, including the inequalities that characterize it. Among these measures, those for reinforcing the mechanisms for assessing educational quality stand out: all the states must have clear performance standards and state assessment systems that are aligned with them, and test all students in grades four through eight annually in English, mathematics and science.

State participation in NAEP testing is a mandatory condition for access to federal funding to support the educational improvement programs contemplated in the new legislation. Student outcomes on state assessments are the criteria used to determine whether the school has made the *adequate yearly progress* (AYP), necessary for receiving federal support, and it can be closed if it fails to do so. This has led to high-stakes assessments having consequences which will be discussed later.

In addition to the United States, from the beginning of the twenty-first century many countries have launched similar assessment systems, including many in Europe as well as East Asia and the Middle East, particularly in Israel, although Arab countries have also started to implement such systems with support from UNESCO. In Africa, a notable example is the South African Consortium for the Monitoring of Educational Quality. In Latin America, Mexico and Costa Rica began to undertake large-scale assessments of elementary education in the seventies and eighties, but only Chile developed a true assessment system prior to 1990. In the last decade of the twentieth century and the first decade of the twenty-first, almost all countries have proceeded to do so (Martínez Rizo, 2009).

At the regional level, a case in point is the Latin American Laboratory for Assessment of Educational Quality (LLECE, acronym in Spanish) of the UNESCO Regional Office for Latin America and the Caribbean. In 1997 it conducted its first

study in third and fourth grade classrooms with the participation of Argentina, Bolivia, Brazil, Chile, Colombia, Cuba, Dominican Republic, Honduras, Mexico, Paraguay and Venezuela (Martínez Rizo, 2008).

Internationally, in addition to the expansion of the IEA assessments, the tests of the Organization for Economic Cooperation and Development (OECD) have been extended further still, through the Programme for Institutional Student Assessment (PISA). These assessments, with a non-curricular approach aimed at fifteen-year old students, were carried out for the first time in 32 countries in the year 2000, and after that, every three years. In 2009 more than 60 countries participated (Martínez Rizo, 2008).

In Mexico, standardized testing began in the second half of the twentieth century with entrance exams for higher education and, at the lower educational levels, with rudimentary tests developed by the teachers themselves or, more often, by supervisors who provided them to the schools under their supervision. Starting in the seventies, the Ministry of Public Education began large-scale assessments. The first tests were administered in 1972, to determine the admission of students to middle school. Towards the end of that decade, the first implementation of assessments in samples of elementary students took place, with a project called *Assessing the academic performance of fourth and fifth graders of elementary school* (Martínez Rizo, 2008).

The situation did not advance much until the early nineties when large-scale assessments received a major boost through the conjunction of several circumstances. The most important one occurred in 1992 with the National Agreement for the Modernization of Elementary Education, which led to the decentralization of the education system and the creation of the *Carrera Magisterial* (Teaching Career), a national teacher incentive program. In order to allocate the salary bonuses that serve as the program's incentives, student outcomes—among other things—are taken into account, which resulted in the need to administer tests to a great many students every year. The first assessment involved more than four million students. These assessments continued until 2005, when the number of students tested reached nearly eight million. A second circumstance emerged with the compensatory programs implemented by the Mexican government with World Bank support. These programs included an assessment component, with testing of the students who were recipients of the program's benefits. Starting in 1994, the implementation of similar tests was extended to include all states on a permanent basis through a project called Study of Elementary Educational Assessment (EVEP, acronym in Spanish).

Also in 1994, Mexico's accession to OECD demonstrated the interest of the authorities in integrating the country into international economic and political life, including participation in educational assessments, such as IEA's TIMSS, LLECE (Latin American Laboratory for the Assessment of Quality in Education) and the PISA project of OECD. In 1996 the task of defining curriculum standards was undertaken

and, in 1998, the National Standards Tests—assessments developed in relation to them—were administered for the first time.

In the twenty-first century, educational assessment in Mexico forged ahead with the creation of the National Institute for Educational Assessment (INEE) and with the development of new initiatives by the Ministry of Public Education (SEP), particularly the census tests called National Assessments of Academic Achievement in Schools (ENLACE, acronym in Spanish) (Martínez Rizo, 2008).

III. Negative consequences and criticism of testing

With the exception of university entrance exams, the tests used at pre-college levels in most U.S. states throughout the twentieth century were low-stakes assessments: their results did not influence important decisions being made either in relation to each student or with respect to the teachers and individual schools. This situation began to change in the eighties, with the trend becoming more pronounced in the nineties, culminating in the provisions of the *No Child Left Behind Act of 2002,* with which large-scale assessments have acquired a fundamental and unprecedented weight in decisions relating to students, teachers and schools.

Something similar has taken place in other countries. The fact that tests are administered on a large scale and their results disseminated through a simple ranking of schools based on students' average scores, regardless of the context in which each one operates, has rendered the results high-stake. This is true even in the absence of specific legal provisions, which involve strong official consequences based on these results, as has happened in the United States, the United Kingdom or Chile.

Much of the criticism of large-scale assessments comes from people who reject them en bloc, without taking into consideration their subtle differences and how their results will be used. On the other hand, the criticism found in the paragraphs below comes from individuals who are knowledgeable about the relevant methodological aspects of standardized testing and, in general, favor proper use of them. Unlike more radical critics, what these appraisals question are what they view as the improper use of assessments, which doesn't take into account the scope and limitations of the tests. Hence, their results tend to be misused, with negative consequences that can be serious (Martínez Rizo, in the press).

Regarding the growing importance that assessment based on testing had acquired in the United States before *No Child Left Behind*, and the concomitant risks it incurred, a recognized expert has said that the trend was due to the—often well-founded—concern of many people with respect to the quality of schools; in this context assessment acquired great influence. He then referred to the negative consequences that resulted from the excessive and misplaced importance that was being given to performance tests:

> Because of the misuse of traditionally constructed standardized achievement tests to judge the quality of schooling, there's some really terrible things happening to our children in schools these days. One of those is important curriculum content is being driven out, because it isn't measured by the test. Another is that kids are being drilled relentlessly on the content of these high-stakes tests and, as a consequence, are beginning to hate school. And a third is that, in many instances, teachers are engaging in test preparation, which hovers very close to cheating, because they're boosting kids' scores without boosting kids' mastery of whatever the test was supposed to measure (Popham, 2001, Secc. Do you think the politicians know this?, par. 2).

Popham (2001) made clear that his position did not refer to all uses of testing for assessing educational outcomes, but rather to inappropriate ways of doing so. He explicitly stated that well-designed, properly used tests can be of great value to education.

> There's a resistance emerging in our country to high-stakes tests of any sort. I think that's unsound. I believe that properly constructed high-stakes tests, tests that can help teachers teach more effectively, should be used. I think the public has a right to know how well their schools are doing. So to resist any kind of testing, I think is disadvantaging the children. You have to create the right kinds of tests. But they can be a powerful force for instructional design, for getting kids to learn what they ought to learn (Popham, 2001, Secc. I met this teacher…, par. 1).

In a recent text, faced with evidence that his fears regarding the spread of large-scale testing without due consideration had become reality, this expert specified two reasons why a good idea—to get students to achieve high levels of competency through a standards-based education—is having the consequences he anticipated: on the one hand, an excess of contents that results in an inadequate definition of standards; on the other, the use of inappropriate tests, particularly instructionally insensitive tests, as a means of verifying achievement of standards (Popham, 2008).

After being signed into law, the experience with the implementation of the *No Child Left Behind Act* has revealed significant shortcomings and counterproductive consequences, especially for public schools. Several projections indicate that very few of these will be able to meet the adequate yearly progress requirements established by the Act, while the great majority (perhaps more than 95% in the entire country) should be classified as failing and forced to face the ensuing consequences, which theoretically might even imply their being "taken over" or "reconstituted" (Oakes and Lipton, 2007).

A leading researcher in the current psychometric field, Robert Linn, before *No Child Left Behind*, also wrote:

> I am led to conclude that, in most cases, the instruments and technology have not been up to the demands that have been placed on them by high-stakes accountability. Assessment systems that are useful monitors lose much of their dependability and credibility for that purpose when high-stakes are attached to

> them. The unintended negative effects of high-stakes accountability uses often outweigh the intended positive effects (Linn, 2000, p. 16).

In the following quote, contemporaneous to the *No Child Left Behind Act,* three scholars from the Rand Corporation explain what is probably behind the high expectations that have led to the misuse of large-scale high-stakes assessments:

> Test-based accountability systems are based on the belief that public education can be improved through a simple strategy: require all students to take standardized achievement tests and attach high stakes to the tests in the form of rewards when test scores improve and sanctions when they do not. (Hamilton, Stecher and Klein, 2002, p. iii).

Many people are unaware of the difficulties involved in achieving good educational outcomes in groups of socially disadvantaged students. In Mexico it is common for business leaders to view the simplistic strategies alluded to in the passage quoted above sympathetically, thinking that the shortcomings of public schools could be easily remedied with private schools such as those attended by their own children; they fail to realize that less than 10% of Mexican children—those from privileged backgrounds—attend them. This would likely explain the widespread view that it would be enough to just apply large-scale testing followed by simple corrective measures in order to substantially improve the quality of education.

In Latin America, until the mid-1990s, the results of large-scale assessments at the primary level did not lead to decisions that could affect individuals, such as deciding whether to pass or fail a student, allocate incentives or take corrective measures that would affect teachers or schools. They were low-stakes or even "no-stakes" assessments because of the lack of widespread circulation of their results. The exception was the System for the Measurement of Educational Quality (SIMCE, its Spanish acronym) in Chile, which from its inception was defined as high-stakes: its census design was expressly conceived in order to contribute to the introduction of major changes in the educational system, such as its decentralization and privatization. The results have been used to determine which schools may receive public funds in the form of individual vouchers for their students.

Recent developments in our *subcontinent* point in a similar direction to that observed in the United States: there is a tendency to think that applying census tests, whose results allow simple and direct comparisons between schools, will facilitate decisions that will lead to significant improvements in the short term. In addition to Chile, Uruguay and Mexico, other countries that are making inroads in the use of census tests are Brazil, Colombia, Costa Rica, Dominican Republic, Ecuador, El Salvador, Guatemala and Peru. The risk of counterproductive consequences is no longer just theoretical, but real.

In Mexico the results of assessment are mixed. On the positive side, technical progress has been achieved and high level specialists have been trained; there is growing public awareness of the right to know the results of assessments, in

contrast to the previous hermetism; and at the federal level and in a few states some education authorities have begun to make use of the results of assessments for decision-making. On the negative side we need to take into account the already excessive and yet still growing number of tests that are developed and implemented and that increasingly burden students, teachers and schools; the prevalence of large-scale assessments that must be used by teachers in the classroom; the increasingly frequent misuse of test results and their excessive influence in the design of public policies.

In short, as large-scale assessments have proliferated, the desire to use their results in support of decisions from which significant improvements in educational quality may be derived has grown. This trend is related to accountability, and acquires significance in the context of broader tendencies: the search for transparency in the management of public affairs; the frequent distrust of public education and, in general, of government-provided services. To this must be added a lack of knowledge of educational testing, not only on the part of the general public, but also among teachers and educational authorities and even among researchers and specialists. This leads to an expectation of almost miraculous results in schools through the use of assessments, regardless of their scope and limits.

In order for the favorable prospects that are associated with assessments to become reality, a fuller picture of their possibilities is needed. We must keep in mind that large-scale assessments have characteristics that limit their ability to inform us about many important aspects that should be included in the curriculum. This is greatly exacerbated in the case of census tests, especially if the aim is to cover many grades with relative frequency. Moreover, such tests can never replace the work of the teacher.

This last point is of particular relevance: only a good teacher can carry out the most important assessment of each student, an evaluation that includes all aspects of the curriculum as well as the more complex cognitive levels, that takes into account each child's circumstances and is repeated with sufficient frequency so as to provide timely feedback, thereby enabling the student to improve. This type of testing should take place regularly in the classroom, with more refined approaches than can be used on a large scale. Many teachers lack the necessary preparation to carry out such an evaluation well, but no large-scale assessment can take its place. Hence, teachers must be given the necessary support to be able to adequately fulfill their evaluative role, and tests could be seen as one of these supports.

Even if done well, the assessments carried out by teachers also have limitations. In particular, their results are not aggregable, in the sense of enabling the construction of synthetic measures nor can they provide information on the status of large-scale contexts, such as educational systems. Large-scale testing can provide valuable input for decision-making at various levels of the system, but only

if it is seen as complementing the work of teachers and is not intended to replace it.

Furthermore, the adoption an attitude that does not perceive assessment as a threat, but as an opportunity for learning and for improvement is essential. The results of assessments—rather than being used to construct simple rankings—could assist in the early detection of at-risk students and schools that need special support and thus provide both with timely assistance, instead of fostering sterile competitions resulting in adverse consequences.

IV. Formative assessment and classroom assessment

If one considers that the tests given by teachers at the end of the nineteenth and beginning of the twentieth centuries consisted, in many cases, of mechanical or rote recitations with which a student demonstrated his knowledge (Oakes and Lipton, 2007), it is not surprising that standardized tests were seen as a step forward and became the benchmark that teachers tried to imitate and for which they were trained.

However, apart from their advantages, large-scale norm-based assessments and multiple choice questions also have clear limitations, especially in relation to the measurement of complex cognitive levels and the difficulty in controlling for the influence of students' social context in the results—in other words, instructional insensitivity. As a result, starting from the very first decades of the last century, there was much criticism of these tests, a questioning which has intensified as large-scale assessments have acquired greater weight, as has occurred in recent decades.

This section will examine the development of an alternative to large-scale assessments, i.e. evaluations performed by teachers. The position that sees large-scale assessments as a complement of teaching work, rather than as a substitute for it, is based on the idea that the influence of a good teacher is irreplaceable, both for students to be able to learn, as well as for assessing the extent to which this occurs—in other words, to evaluate.

Assessing the degree to which a student has acquired the knowledge and skills expected at the end of a school year is not easy if one wishes to adequately cover the different subjects or areas of the curriculum and the topics included in each subject. The task becomes more complicated if one also wishes to know the progress made by each student—which is essential for providing feedback—since such assessments must be made at the start of the school year and at various times throughout the year, on a permanent basis.

This last is essential if we wish the evaluation to be useful not only for determining the outcome of an educational process—what is referred to as a summative evaluation—but, above all, for contributing to an improvement in the learning process as a whole—the formative evaluation. If we're assessing the daily

progress of twenty or thirty students, and we want to have information about their personal, family and social circumstances in order to take it into account when making important decisions for the future of each of them, the task of evaluating becomes complex.

As previously mentioned, from the early development of large-scale testing, some of its most lucid promoters, such as Tyler, pointed out that this type of assessments should also be seen as part of the teaching-learning process, but the approach that prevailed actually treated them as an additional element that just took place at the end of the process.

The distinction between the final assessment and that which takes place throughout the process, between *formative* and *summative assessment*, is recent.

Classical Test Theory and traditional design large-scale performance assessments were developed during the first half of the twentieth century; both were marked by the psychological and pedagogical theories of the time, prominent among which are schools of thought such as Skinner's behaviorism. From mid-century on, the development of new psychometric conceptions occurred concurrently with the so-called *cognitive revolution*, from which pedagogical trends that fall under the overused label of *constructivism* were also derived. These developments coincided in rejecting the behaviorist approach that reduces the field of psychology to the study of the most directly observable phenomena, and instead attempted to open the black box of the mind and explore the processes that take place inside, through such techniques as *thinking aloud* (Shepard, 2006).

To the extent that mental processes are identified and explored, vast and attractive horizons open up both for teaching as well as for learning assessment methodologies—particularly for those who intend to pursue work in *formative assessment*—by providing elements that enable teachers and students to modify their actions, thereby achieving better results. In this regard, one notable paper on classroom assessment points out important elements in reference to the formative potential of assessment:

> Assessment cannot promote learning if it is based on tasks or questions that divert attention from the real goals of instruction. Historically, traditional tests have often misdirected instruction, if they focused on what was easiest to measure instead of what was important to learn. (Shepard, 2006, p. 626).

Interest in classroom assessment for formative purposes is related to increasing awareness of the limitations of conventional testing for such ends, and to the parallel progress made by experts in the area of curricular content. These experts—both because of a rejection of the effects of assessments used for accountability as well as because of profound changes in the conceptions of learning and appropriate content management—began to develop alternatives to these assessments for classroom use (Shepard, 2006).

As already mentioned, many teachers lack the competence to manage classroom assessments that are superior to large-scale assessments in terms of their potential feedback on the work of teachers and that of their students. For this reason, as far back as 1989, Silver and Kilpatrick (as quoted in Shepard, 2006, p. 627) argued that:

> Rather than the prevailing practice wherein teachers oriented their own tests to emulate both the form and content of external, multiple-choice tests, a serious effort should be made to *reskill* teachers to conduct problem-solving lessons and to assess their students' problem-solving abilities and dispositions in the context of those lessons.

A more recent paper presents an interesting summary of the way in which the definition of formative assessment has evolved (Brookhart, 2009).

The original idea that distinguishes between the information that is used to improve a work in process, as opposed to that which is used to assess the final result, was proposed by Michael Scriven in 1967, in reference to the evaluation of curriculum and educational programs. Other authors soon came to realize the importance of this distinction, which—although it now seems obvious—had not been explicitly established prior to Scriven's seminal work.

In 1971 Bloom, Hasting and Madaus' book appeared, popularizing the concepts of formative and summative assessment, as applied to students' learning. This work identifies the differences found in assessments used to support instructional decisions, distinguishing between formative and summative purposes, as well as location and diagnosis. Brookhart (2009) emphasizes that this work adds an important element to Scriven's concept: in addition to *providing information about the learning process as opposed to just its final results*, formative assessment also provides information *that can help teachers make better instructional decisions*. We might add that Bloom put his ideas into practice with the teaching system Mastery Learning, based on Carroll's model of learning.

The concept was further developed with Sadler (1989), who considered that not only were the results of formative assessments useful for teachers, but that students could use them as well. This author applies the adjective *formative* to the noun used to refer to the evaluation of students' learning—*assessment*—rather than applying it to the word *evaluation*, as in Scriven and Bloom, where it refers to curricula and programs (Brookhart, 2009).

Another step forward in refining the idea of formative assessment occurs with the emphasis given to the importance of affective aspects in the feedback given to students, in contrast to the previous emphasis on cognitive aspects. More recently, authors such as Black and Wiliam (1998), Stiggins (2008) and Brookhart (2009), have called attention to this dimension.

The last author cited notes that, until recently, it was considered acceptable if only a few students achieved learning objectives, even if many more failed to do so.

The role of testing was to distinguish the first group of students from the second, and the criteria for evaluating the quality of the assessments were their validity and reliability. Today schools are expected to succeed in getting all students to attain the necessary levels of competency, compelling us to reflect on appropriate ways to assess learning in this new context, which necessarily involves the emotional impact of assessment on students. Stiggins states:

> From the very earliest grades, some students… scored high on assessments and were assigned high grades. The emotional effect of this was that they came to see themselves as capable learners—they became increasingly confident in school… But other students scored very low on tests right from the beginning and so they were assigned failing grades. This caused them to doubt their own capabilities as learners from the outset. Their loss of confidence deprived them of the emotional reserves to continue to risk trying… If a student gave up and stopped trying (even dropped out of school), it was regarded as that student's problem, not the teacher's or school's problem. The school's responsibility was to provide the opportunity to learn. If students didn't take advantage of the opportunity, that was not the system's responsibility (2008, p. 7).

Stiggins (2008) adds that the importance of the paradigm shift, which involves focusing attention on the students as privileged users of outcomes, taking into account the emotional impact of assessments, cannot be overstated:

> Over the decades, school improvement experts have made the mistake of believing that the adults in the system are the most important assessment users… We have believed that, as the adults make better instructional decisions, schools will become more effective… But this perspective overlooks the reality that students may be even more important data-based instructional decision makers than the adults… If a student decides that the learning is beyond reach for her or him or that the risk of public failure is too great and too embarrassing, then regardless of what we adults do, the learning stops. So the essential question for teachers and school leaders is: What can we do to help students answer the above questions in productive ways that keep them believing that success is within reach for them if they keep trying? (p. 8).

Each of the stages in the development of the concept of formative assessment has contributed something substantive: Scriven's original idea, which distinguishes between the assessment taking place during or at the end of a process; the explicit application of the concept to the assessment of learning, not just to curriculum or programs, as in Bloom; the identification of students as key recipients of the information, with Sadler; and, finally, the focus on the emotional dimension, with Brookhart, Black and Wiliam and Stiggins.

Conclusion

Formative assessment, in the classroom or on a broader level, is not easy, but if we fail to adopt such an approach, the usefulness of assessment as a tool for improvement will be reduced. Therefore, providing teachers with the elements

necessary for orienting their classroom assessments toward a formative approach is important as well as complex.

It would seem that the trajectory that assessment is currently following in our educational system is not headed in the right direction. While acknowledging the positive aspects of developments that have occurred in recent years in Mexico in regard to educational assessment, it appears that the time has come to issue a warning. This wake-up call might be more productive if, at the same time, we propose a better alternative. The alternate course of action is none other than an assessment system that combines—in a more balanced manner—sparse and consistent large-scale assessment with rich formative assessment work conducted in the classroom by teachers.

This idea is developed in these last few paragraphs, again following Stiggins' (2008) text, titled—significantly—*Assessment Manifesto: A Call for the Development of Balanced Assessment Systems*.

> A *manifesto* is a public statement of intention, belief, opinion, or policy advocating political and/or social action. Often such ardent statements run counter to conventional or dominant values and practices within the context in which they are issued. I issue this assessment manifesto because I believe that we have reached a tipping point in the evolution of our schools when we must fundamentally reevaluate, redefine, and redesign assessment's role in the development of effective schools. The work to be done is so crucial as to require urgent pedagogical, social, and political action. (p. 2).

In the last pages of Stiggins' (2008) text, the author explains what his manifesto consists of, in terms of a total assessment solution:

> We understand far more today than ever before about how to use assessment productively. We must replace grossly out-of-balance assessment systems of the past with those that honor the information needs of all assessment users—systems that both support and verify learning from the classroom to the boardroom. To attain long-missing and much-needed balance, we must implement classroom assessment practices that rely on an ongoing array of quality assessments used strategically in ways that keep students believing in themselves… It is time to replace the intimidation of accountability as our prime motivator with the promise of academic success for all learners as that motivational force. It is not that intimidation is universally ineffective, but it only motivates those who have hope of success. *Unfortunately, true hopelessness always trumps intimidation when it comes to learning*. Effective classroom assessment can and must serve to promote hope in all students (p. 10).

Stiggins (2008) notes that today we have the necessary conditions to modify assessment systems and point them in the right direction; that, thanks to research that has been conducted over the last two decades we are in a position to implement training activities that provide teachers with the competencies

necessary to assess effectively, and he argues that schools of education should pursue such a course.

The situation of the Mexican educational system is similar, but more serious. Hence the need to balance our assessment system is even more pressing. Stiggins ends his manifesto by stating that we have what is needed and the only question that needs to be answered is one that should also be posed in Mexico:

> *Will practitioners and policymakers be given the opportunity to learn to assess productively*? Historically, the answer has been an unequivocal, "No, they will not." As a result, the immense potential of assessment to support student learning has gone untapped—indeed, unnoticed at the highest levels of policy making. It need not be so. We have in hand a new vision of excellence in assessment that will tap the wellspring of confidence, motivation, and learning potential that resides within every student. The time has come to embrace it (Stiggins, 2008, p. 12).

A review of the literature on the subject of formative classroom assessment reveals the growing interest it has aroused in the educational media. References were rare in the 1980s; they increased throughout the 1990s, particularly in the second half of the decade; and they have become numerous in the twenty-first century. An as yet unpublished paper with two hundred references on the subject, most of them from the last decade, is available to anyone who might be interested (Martínez Rizo, 2009).

References

Baker E. L. & Choppin, B. H. (1990). Minimum competency testing. In H.J Walberg & H. J. Haertel (Eds.). (1990). *The International Encyclopedia of Educational Evaluation* (pp. 499-502). Oxford-New York: Pergamon Press.

Black, P. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*, 7-74.

Bloom, B. S, Hastings, J. T., & Madaus, G. F. (Eds.). (1971). *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill.

Brookhart, S. M. (2009). Editorial. *Educational Measurement: Issues and Practice, 28* (1), 1-2.

De Landsheere, G. (1996). *La investigación educativa en el mundo* (Trans. G. A. Gallardo Jordán). Mexico: Fondo de Cultura Económica. (Original work published in 1986).

Hamilton, L. S., Stecher, B. M., & Klein S. P. (Eds.). (2002). *Making sense of test-based accountability in education*. Santa Monica, CA: Rand Corporation.

Klein, S. P. & Hamilton, L. (1999). *Large-scale testing. Current practices and new directions.* Santa Monica, CA: Rand Education.

Linn, R. (2000). Assessments and accountability. *Educational Researcher, 29* (2), 4-16.

Martínez, Rizo, F. (2008). *Las evaluaciones educativas en América Latina* (Serie: Cuadernos de Investigación, No. 32). Mexico: Instituto Nacional para la Evaluación de la Educación.

Martínez Rizo, F. (2009). *Marco de referencia para el proyecto "El uso formativo de la evaluación para mejorar el aprendizaje"*. Unpublished manuscript, Universidad Autónoma de Aguascalientes-Instituto de Investigación, Innovación y Estudios de Posgrado de la Educación.

Martínez Rizo, F. (in press). Assessment in the context of educational policy: The case of Latin American Countries. In E. Baker, B. McGaw, & P. Paterson (Eds.), *International Encyclopedia of Education* (3rd ed.). Oxford-New York: Elsevier.

Mathison, S. (2008). A short history of educational assessment and standards-based educational reform. In S. Mathison, & E. W. Ross (Eds.), *The nature and limits of standards-based reform and assessment* (pp. 3-14.). New York: Teachers College Press.

Oakes, J. & Lipton, M. (2007). *Teaching to change the world* (3rd ed.). New York: McGraw Hill.

Phelps, R. P. (2000). Trends in large-scale testing outside the United States. *Educational Measurement: Issues and Practice, 19* (1), 11-21.

Popham, W. J. (2001). Interview: James Popham. *Frontline.* Retrieved October 7, 2009, from: http://www.pbs.org/wgbh/pages/frontline/shows/schools/interviews/popham.html

Popham, W. J. (2008). Standards-based education: Two wrongs don't make a right. In S. Mathison & E. W. Ross (Eds.), *The nature and limits of standards-based reform and assessment* (pp. 15-25). New York: Teachers College Press.

Postlethwaite, N. (1985). International association for the evaluation of educational achievement (IEA). In T. Husén & N. Postlethawite (Eds.), *International encyclopedia of education* (pp. 2645-2646). New York: Pergamon Press.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119-144.

Scriven, M. (1967). The methodology of evaluation. In R. Tyler, R. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39-83). Chicago, IL: Rand McNally.

Shepard, L. A. (2006). Classroom assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th. ed., pp. 623-646). Westport, CT, United States: Praeger.

Stiggins, R. (2008). *Assessment manifesto: A call for the development of balanced assessment systems.* Portland, United States: ETS Assessment Training Institute.

Translator: Jeanne Eileen Soennichsen