THE WORLD BANK GROUP

Framework for Building an Effective Student Assessment System

READ/SABER Working Paper

Marguerite Clarke

8/18/2011

Contents

Abstract

The purpose of this paper is to help countries understand some of the *key principles and characteristics of an effective student assessment system*. The focus is on assessment of student learning and achievement at the K-12 level.[1] The paper extracts principles and guidelines from countries' experiences, professional testing standards, and the current research base to provide policy makers, development organization staff, and others with a *framework and key indicators for diagnosis, discussion, and consensus-building around how to construct a sound student assessment system that supports improved quality and student learning*.

[1] The paper does not discuss psychological or workplace testing; nor does it explicitly discuss assessment at the tertiary level (although many of the issues also apply to that level).

"[Assessment] goes to the heart of what matters in education: not just enrollment and completion rates, but the ultimate goal of student learning" (World Bank, 2010, p.5).

Introduction

Assessment is the process[2] of gathering and evaluating information on what students know, understand, and can do in order to make an informed decision about next steps in the educational process. Data collection and evaluation methods can be as simple as oral questioning and response (for example, "What is the capital of Ethiopia?"), or as complex as computer-adaptive testing models based on multifaceted scoring algorithms and learning progressions.[3] Decisions made based on the results may vary from how to design system-wide programs to improve teaching and learning in classrooms, to identifying next steps in classroom instruction, to determining which applicants should be admitted to university.

An *assessment system* is a group of policies, structures, practices, and tools for generating and using information on student learning. Effective assessment systems are those that provide information of sufficient quality and quantity to meet stakeholder information and decision-making needs in support of improved quality and student learning (Ravela et al., 2009).[4]

Governments, international organizations, and other stakeholders are increasingly recognizing the importance of assessment for monitoring and improving student learning, and the

[2] When used as a noun, *assessment* may refer to a particular tool, such as a test.
[3] A list of computer-adaptive testing programs can be found at http://www.psych.umn.edu/psylabs/catcentral/.
[4] A student assessment system supports a variety of information purposes or needs, such as informing learning and instruction, determining progress, measuring achievement, and providing partial accountability information. All of these purposes, and the decisions based on them, should ultimately lead to improved quality and learning levels in the system.

concomitant need to develop strong systems for student assessment (IEG, 2006; McKinsey & Company, 2007; UNESCO, 2007). This recognition is linked to growing evidence that the benefits of education accrue to society only when learning occurs (Hanushek and Woessmann, 2007, 2009; OECD, 2010). For example, a one standard deviation increase in scores on international assessments of reading and mathematics is associated with a 2 percent increase in annual growth rates of GDP per capita.

Some people argue that assessments, particularly large-scale assessment exercises, are too expensive. In fact, the opposite tends to be true, with *testing shown to be among the least expensive innovations in education reform*, costing far less than increasing teachers' salaries or reducing class size. Hoxby (2002) found that even the most expensive state-level, test-based accountability programs in the United States cost less than 0.25 percent of per-pupil spending. Similarly, in none of the Latin American countries reviewed by Wolff (2007) did testing involve more than 0.3 percent of the national education budget at the level (primary or secondary) tested.[5]

Over the last 20 years, many countries have started implementing assessment exercises or building on existing assessment systems (UNESCO, 2007). In addition, there has been huge growth in the number of countries participating in international comparative assessment exercises such as the Trends in International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA).[6] Nongovernmental organizations also have increasingly

[5] Others argue that investing in assessment should be seen in terms of the *use* to be made of the results rather than on the basis of variables like cost per pupil.

[6] For example, the number of countries participating in PISA jumped from 43 in 2000 to 66 in 2007. A comparatively small number of developing countries have participated in international assessments of student

turned to student assessment to draw public attention to poor achievement levels and to create an impetus for change.

Despite this, far too few countries have in place the policies, structures, practices, and tools that constitute an effective assessment system. *This is particularly the case for low-income countries, which stand to benefit most from systematic efforts to measure learning outcomes*. Some of these countries have experimented with large-scale or other standardized assessments of student learning, but too often these have been ad hoc experiences that are not part of an education strategy and are not sustained over time. A key difference between one-off assessments and a sustained assessment system is that the former provides a snapshot of achievement while the latter allows for the possibility of monitoring trends over time (more like a series of photos) and a better understanding of the relative contribution of various inputs and educational practices to changes in those trends. One-off assessments can generate shock value and an opening for discussions about education quality, and this can be a short-term strategy for putting learning on the agenda.[7] Ultimately, however, governments must deal with the challenging, but necessary, task of putting in place systems that allow for regular monitoring of, and support for, student learning. This is the only way to harness the full power of assessment.

---

achievement. These countries have consistently performed in the bottom of the distribution, limiting the amount of information they can derive from the data to better understand and improve their own education systems.

[7] One of the more popular of these initiatives is known as EGRA. According to the USAID Website (https://www.eddataglobal.org/): "The Early Grade Reading Assessment (EGRA) is an oral assessment designed to measure the most basic foundation skills for literacy acquisition in the early grades….in order to inform ministries and donors regarding system needs for improving instruction."

Theory and Evidence on Student Assessment

A basic premise of the research on assessment is that the right kinds of assessment activities, and the right uses of data generated by those activities, contribute to better outcomes, be those improved learning or improved policy decisions (for example, Heubert and Hauser, 1999).[8] What constitutes 'right' is driven by a set of theoretical and technical guidelines for test developers and users of testing information (AERA, APA, and NCME, 1999).

There also is a sizeable body of empirical research showing the benefits of specific types of tests, when implemented and used correctly, on student learning. For example, research demonstrates the link between high-quality, formative classroom assessment activities and better student learning outcomes as measured by performance on standardized tests. Black and Wiliam's (1998) synthesis of over 250 studies from around the world on the impact of effective classroom assessment shows gains of a half to a full standard deviation on standardized tests, with the largest gains being realized by low achievers.[9] These findings have important implications for the closing of achievement gaps among different student groups.

Research on exit examinations demonstrates a link between countries that have those policies and higher performance levels on international assessments, such as PISA or TIMSS (for example, Bishop, Mane, and Bishop, 2001). At the same time, these types of examinations have

[8] Ravela et al. (2008) note that assessment is a necessary, but not sufficient, 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 inservice 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.

[9] Rodriguez (2004) reports effects of similar size in U.S. TIMSS mathematics performance arising from the effective management of classroom assessment (this findings is based on analysis of the responses of teachers from TIMSS participating countries to questions on the topic of management of classroom assessment).

been shown to have a negative impact on students from disadvantaged groups by limiting their opportunities to proceed to the next level of the education system or avail themselves of certain kinds of educational opportunities (Greaney and Kellaghan, 1995; Madaus and Clarke, 2001). Because of this, the uses and outcomes of examinations must be carefully monitored at the system, group, and individual levels, and efforts made to reduce or mitigate any unintended negative consequences.

Research shows a weak but positive link between the uses of data from large-scale assessments to hold accountable schools and educators and better student learning outcomes (for example, Carnoy and Loeb, 2002). Key determinants of whether the effects are more positive than negative appear to be the degree of alignment between test design and test score use and the extent to which supports are in place to help struggling or underperforming schools or teachers (Ravela, 2005).

Research is increasingly focusing on the characteristics of effective assessment *systems* that encompass the aforementioned types of assessment activities and uses (that is, classroom assessment, examinations, and large-scale assessment). This research draws on the principles and best practices found in the assessment literature as well as analyses of the assessment systems of high-achieving nations. Darling-Hammond and Wentworth (2010) reviewed the practices of high-performing systems around the world (for example, Australia, Finland, Singapore, Sweden, and the United Kingdom), and noted that these systems:

- illustrate the importance of assessment *of*, *for*, and *as* learning, rather than as a separate disjointed element of the education enterprise;

- provide *feedback* to students, teachers, and schools about what has been learned, and *'feed forward'* information that can shape future learning as well as guide college- and career-related decision making;
- closely *align* curriculum expectations, subject and performance criteria, and desired learning outcomes;
- *engage teachers* in assessment development and scoring as a way to improve their professional practice and their capacity to support student learning and achievement;
- *engage students* in authentic assessments to improve their motivation and learning;
- seek to advance student learning in *higher-order thinking skills and problem solving* by using a wider range of instructional and assessment strategies;
- privilege *quality over quantity* of standardized testing;[10] and
- as a large and increasing part of their examination systems, use *open-ended performance tasks and school-based assessments* that require students to write extensively and give them opportunities to develop 'twenty-first century' skills.[11]

While this research tells us what an effective system looks like, it does not tell us what it takes to get there. Other studies delve into these capacity-building and strategy issues. For example, Ferrer (2006) provides advice on designing sustainable and sound assessment systems based on his analysis of existing systems in Latin America. Bray and Steward (1998) carry out a similar

[10] That is to say, some countries have good learning outcomes, but don't test a lot (for example, Finland). Other countries test a lot (for example, the United States), but don't do so well on international assessments.

[11] Standardized performance tasks are incorporated into examination scores in systems as wide-ranging as the GCSE in the United Kingdom, the Singapore examinations system, the certification systems in Victoria and Queensland, Australia, and the International Baccalaureate, which operates in more than 100 countries around the world. Because these assessments are embedded in the curriculum, they influence the day-to-day work of teaching and learning, focusing it on the use of knowledge to solve problems.

analysis for secondary school examinations. Others (for example, Lockheed, 2009) evaluate the status of donor activity in these areas and discuss how to improve the effectiveness of this support to countries.

This paper draws together these streams of evidence to create a unified framework for understanding *what* an effective student assessment system looks like and *how* countries can begin to build such systems.

Framework for Student Assessment Systems

In order to approach this framework in a strategic way, we need to identify some key dimensions of effective assessment systems. Two main dimensions are discussed here: (i) *types/purposes* of assessment activities and (ii) the *quality* of those activities.

Dimension I. Assessment Types/Purposes

Assessment systems tend to comprise three main kinds of assessment activities, corresponding to three main information needs or purposes (see also Annex 1). These kinds and the concomitant information needs/purposes are:

- *classroom assessments* for providing real-time information to support teaching and learning in individual classrooms;
- *examinations* for making decisions about individual student's progress through the education system (for example, certification or selection), including the allocation of 'scarce' educational opportunities; and

- *large-scale assessments* for monitoring and providing policy- and practitioner-relevant information on overall performance levels in the system, changes in those levels, and related or contributing factors.

To be sure, these assessment types/purposes are not completely independent of each other; nor are they all-encompassing (that is, there are some assessment activities that don't quite fit under these labels). At the same time, they represent the main kinds of assessment activities carried out in the majority of education systems around the world.

Classroom assessments, also referred to as continuous assessments, are those carried out by teachers and students in the course of daily activity (Airasian and Russell, 2007). They encompass a variety of standardized and non-standardized instruments and procedures for collecting and interpreting written, oral, and other forms of evidence on student learning or achievement. Examples include oral questioning and feedback, homework assignments, student presentations, diagnostic tests, and end-of-unit quizzes. The main purpose of these assessments is to provide 'real time' information to support teaching and learning. They encompass assessment *for* learning (that is, determining the next step in the instructional process based on what the student already knows and can do) and assessment *as* learning (helping students to reflect on evidence of learning so that they become more aware of what they learn, how they learn, and what helps them learn).

Examinations, variously modified by the terms 'public,' 'external,' or 'end of cycle,' provide information for high-stakes decision making about individual students—for example, whether they should be assigned to a particular type of school or academic program, graduate from high

school, or gain admission to university (Greaney and Kellaghan, 1995; Heubert and Hauser, 1999). Whether externally administered or (increasingly) school-based, their typically standardized nature is meant to ensure that all students are given an equal opportunity to show what they know and can do in relation to an official curriculum or other identified body of knowledge and skills (Madaus and Clarke, 2001). The leaving certificate examinations at the end of compulsory education in many education systems are a good example. The high-stakes nature of most examinations means they exert a backwash effect on the education system in terms of what is taught (resulting in "teaching to the test" or even "teaching the test") and learned, having an impact, for better or worse, on the skills and knowledge profile of graduates (West and Crighton, 1999). These tests have potentially negative consequences for individual students, particularly those from disadvantaged groups, who may be excluded from the education of their choice (or any kind of education at all) on the basis of their performance (Greaney and Kellaghan, 1995). Such consequences must be considered when determining whether the use of such tests is appropriate.[12] It is important to emphasize that there are very specific professional and technical standards regarding the appropriate and inappropriate uses of examinations (and tests in general) for making high-stakes decisions about individual students (AERA, APA, and NCME, 1999).

[12] Greaney and Kellaghan (1995) note that because of the high stakes attached to examination performance, teachers often teach to the examination, with the result that inadequate opportunities to acquire relevant knowledge and skills are provided for students who will leave school at an early stage. Practices associated with examinations that may create inequities for some students include scoring practices, the requirement that candidates pay fees, private tutoring, examination in a language with which students are not familiar, and a variety of malpractices. The use of quota systems to deal with differences in performance associated with location, ethnicity, or language-group membership also creates inequities for some students.

Large-scale assessments are designed to provide information on system performance levels and related or contributing factors (Greaney and Kellaghan, 2008; Kifer, 2001), typically in relation to an agreed-upon set of standards or learning goals, in order to inform educational policy and practice. Examples include international assessments of student achievement levels such as TIMSS, PIRLS, and PISA; regional assessments such as PASEC in Francophone Africa, SACMEQ in Anglophone Africa, and LLECE in South America; national-level assessments such as SIMCE in Chile; and subnational assessments such as state-level tests in the United States or Canada.[13] These assessments vary in the grades or age levels tested, coverage of the target population (sample or census), subjects or skill areas covered, types of background data gathered, and the frequency with which they are administered. They also vary in how the results are reported and used. For example, while some stop at the reporting of results to policy makers or the general public, others use the results to hold accountable specific groups in the education system (Clarke, 2007). Ravela (2005) describes the use of large-scale national assessment results in Uruguay to help teachers improve their teaching. The emphasis on formative uses at the classroom level helped enhance teacher acceptance of the results; it also influenced the assessment design in terms of the need to use a census-based approach to data collection and the use of background factors to control for non-school factors affecting achievement.[14]

[13] TIMSS – Trends in International Mathematics and Science Study; PIRLS – Progress in International Reading Literacy Study; PISA – Program for International Student Assessment; PASEC – Programme d'Analyse des Systèmes Educatifs (Program on the Analysis of Education Systems); SACMEQ – Southern and Eastern Africa Consortium for Monitoring Educational Quality; LLECE – Latin American Laboratory for Assessment of the Quality of Education; Sistema de Medición de Calidad de la Educación.

[14] World Bank support for assessment activity over the last 20 years (Larch and Lockheed, 1992; Liberman and Clarke, 2011) has shifted from an emphasis on examination reform to an emphasis on the implementation of large-scale assessment exercises to monitor achievement trends and inform policy and practice.

One way to differentiate among the above three types of assessment activities is that classroom assessment is mainly about assessment *as* learning or *for* learning (and hence is primarily formative in nature) while examinations and surveys are mainly about assessment *of* learning (and hence are primarily summative in nature). These distinctions do not always hold up neatly in practice and hybrid approaches are becoming more common. For example, Singapore has an assessment system structured around public examinations, but has built a whole infrastructure of support *for* learning around it (L. Benveniste, personal communication, March 2010). Other hybrid activities involve the adaptation of tools designed for one type of assessment activity (for example, classroom instruments for informing instruction) for another purpose (for example, documenting performance at the system level). One of the best known of these initiatives is Early Grade Reading Assessment, or EGRA for short, an instrument developed with the support of donor agencies and experts for use in developing countries (https://www.eddataglobal.org/). Based on a tool originally designed for classroom use, EGRA has been used to collect system-level data on student performance on early reading skills in order to inform ministries and donors regarding system needs for improving instruction.

Education systems can have very different profiles in these three assessment areas insofar as their purposes and related uses for assessment vary. There is no one ideal profile. For example, Finland's education system emphasizes classroom assessment as a key source of information on student learning and draws less on examinations or large-scale assessment. China has traditionally placed considerable emphasis on examinations as a means to sort and select from

its large student population, and relatively less on classroom assessment or large-scale survey exercises (although this is changing).[15]

Dimension 2. Quality Drivers

Instead of being able to reference one ideal 'profile' for an effective assessment system, the key consideration is the individual and combined quality of the assessment activities in terms of the adequacy of the information generated to support decision making (Messick, 1989; Shepard, 2000).

There are three main drivers of information quality in an assessment system (AERA, APA, and NCME, 1999; Darling-Hammond and Wentworth, 2010):

- *enabling context*,
- *system alignment*, and
- *assessment quality*.

Although closely related, these dimensions are presented here separately for the purposes of elucidation and discussion.

The *enabling context* refers to the broader context in which the assessment activity takes place and the extent to which that context is conducive to, or supportive of, the assessment. It covers such areas as the broader legislative or policy framework for assessment activities; the

[15] Several factors contribute to countries' differing profiles in relation to these assessment activities. One important contributing factor is the official vision and goals for the education system and the perceived role of assessment in achieving that vision. Another is the historical legacy of assessment in a particular education system, which can create a pull toward a particular type of assessment activity (Madaus, Clarke, and O'Leary, 2003). Still another is the capacity of various stakeholders in the system to effectively carry out different types of assessment activities (Greaney and Kellaghan, 2008). Yet another is the cost, perceived or real, of assessment activities (Wolff, 2007).

institutional and organizational structures for designing, carrying out, or using the results from the assessment activity;[16] the availability of sufficient and stable sources of funding; and the presence of competent assessment unit staff and classroom teachers. The enabling context is important to get right because it is a key driver of the long-term quality and effectiveness of an assessment system and—like the soil, water, and air that a plant needs to grow—no assessment system is sustainable in its absence (World Bank, 2010). In most instances, the onus is on the government to at least provide the vision, leadership, and policy framework toward establishing this enabling context, which may subsequently be implemented via public-private partnerships. Some education systems, particularly in federal contexts, combine forces to create an enabling context in terms of pooling resources or institutional arrangements for developing, implementing, analyzing, or reporting on tests. Regional assessment exercises, such as SACMEQ, PASEC, and LLECE, represent another form of collaboration toward creating an enabling context. The efficiencies of scale achieved by these collaborations make it more cost effective to develop higher-quality tests and to incorporate technological advances into the testing process.

System alignment refers to the extent to which the assessment system is aligned with the rest of the education system. This includes the connection between assessment activities and system learning goals, standards, curriculum, and pre- and in-service teacher training opportunities (Fuhrman and Elmore, 1994; Smith and O'Day, 1991). It is important for assessment activities to align with the rest of the education system so that the information they

[16] There is much debate over whether assessment units should be located within or outside of education ministries. In fact, the institutional location is not as important as the culture of continuity and transparency created around assessment (Ravela et al., 2008). Such a culture is achieved when an assessment has a clear mandate and solid structure, which necessitates that the assessment system be underpinned by some kind of legal statute.

provide is of use to improving the quality of education in the system, and so that synergies can be created. Alignment considerations for assessment systems include:

- domain coverage—the extent to which assessment activities provide information on student learning and achievement in relation to the curriculum in general and key knowledge, skills, and competencies in particular;
- population/system coverage—the extent to which assessment activities provide information on all students at all grades; and
- utility—the extent to which assessment activities are consistent with, and useful/usable in relation to, stakeholder learning goals and priorities.

It is evident that alignment involves more than the simple match between what is tested and what is in the curriculum. Hence, while the correspondence between a country's curriculum and what is tested on international assessments such as PISA and TIMSS may be low, the assessment might still be aligned with (and useful for informing) the overall goals of the education system and any related reforms underway or planned. Indeed, the use of data from TIMSS, PIRLS, and PISA to identify drivers of performance and monitor the impact of reforms on performance over time has been key to the improvement of achievement levels in countries as diverse as Brazil, Jordan, and Poland.

Assessment quality refers to the psychometric quality of the instruments, processes, and procedures used for the assessment activity (AERA, APA, and NCME, 1999). It is important to note that assessment quality is a concern for *any kind of assessment activity* – that is, classroom assessment, examinations, or large-scale assessment. It covers such issues as the *design and implementation* of assessment activities, examination questions, or survey items; the *analysis*

and interpretation of student responses to those assessment activities, questions, or items; and the appropriateness of how the assessment, examination, or survey results are *reported and used* (Heubert and Hauser, 1999; Shepard, 2000). Depending on the assessment activity, the exact criteria used to make those judgments differ. Assessment quality is important because if an assessment is not sound in terms of its design, implementation, analysis, interpretation, reporting, or use, it may contribute to poor decision-making in regards to student learning and system quality (Messick, 1989; Wolff, 2007).

Two overarching technical issues for any assessment are reliability and validity. *Reliability* refers to whether the assessment produces accurate information, and is a particularly important consideration for high-stakes examinations and for monitoring trends over time. *Validity* pertains to whether the test scores represent what they are supposed to represent and whether they can be used in the intended ways. One common threat to test score validity is a difference between the language of instruction and the language of testing, which may make it difficult for a child to show what they know and can do. Use is a very important concept in relation to validity, and requires a careful consideration of the consequences of test score use, including the social, economic, and other impacts on different groups in the population.

Crossing these quality drivers with the different assessment types/purposes, we arrive at the framework shown in Table 1.

Table 1. Framework for Building an Effective Student Assessment System

	Assessment types/purposes		
	Classroom assessment	**Examinations**	**Large-scale assessment**
Enabling context			
System alignment			
Assessment quality			

Source: Author.

The rest of this paper fleshes out and discusses the use of this framework for building a more effective assessment system. The framework can be applied to any country's assessment system as a way to determine where the system is strong and where more work is needed.

Fleshing out the Framework

The framework in Table 1 is a starting point for identifying indicators that can be used to review assessment systems and plan for their improvement. Indicators can be identified based on a combination of criteria, including:

- professional standards for assessment; and
- empirical research on the characteristics of effective assessment systems, including analysis of the characteristics that differentiate between the assessment systems of low- versus high-performing nations.

Where there are no professional standards, or where the empirical research is limited, we can select indicators based on two additional criteria: (i) theory-driven—there is consensus among experts that it contributes to effective assessment; and (ii) resource-driven—a majority of governments make substantial investments in the area.

The evidence base is stronger in some areas than in others. For example, there are many professional standards for assessment quality that can be applied to classroom assessments, examinations, and large-scale assessments (APA, AERA, and NCME, 1999),[17] but less professional or empirical research on enabling contexts.

The above criteria were used to identify the indicator areas shown in Table 2. These indicator areas are most relevant to examinations and large-scale assessment activities, but, with some modifications, also can be applied to classroom assessment.

[17] There also is a sizeable research base on system alignment (for example, Fuhrman and Elmore, 1994; Hamilton, Stecher, and Klein, 2002).

Table 2. Framework for Building an Effective Student Assessment System with Indicator Areas

	Assessment types/purposes		
	Classroom assessment	**Examinations**	**Large-scale assessment**
Enabling context	Policies Fiscal resources Organizational structures Human resources		
System alignment	Learning goals and standards Curriculum Pre- and in-service teacher training		
Assessment quality	Design Administration Analysis Uses		

Source: Author.

Data for some of these indicator areas can be found in official documents, published reports (for example, Ferrer, 2006), research articles (for example, Braun and Kanjee, 2005), and online databases.[18] For the most part,[19] data have not been gathered in any comprehensive or systematic fashion. Those wishing to review this type of information for a particular assessment system will most likely need to collect the data themselves. Standardized questionnaires and rubrics for collecting and evaluating data on each of the three assessment areas (classroom assessments, examinations, and large-scale assessment) are available at

[18] Two of the more useful online databases are http://www.inca.org.uk/ and http://epdc.org/.
[19] Brinkley, Guthrie, and Wyatt (1991) surveyed large-scale assessment and examination practices in OECD countries. Larach and Lockheed (1992) did a similar survey of assessments supported by the World Bank. Macintosh (1994) did a study in 10 countries (Australia, Bahrain, England and Wales, Guatemala, Israel, Malaysia, Namibia, Poland, Scotland, and Slovenia).

http://www.worldbank.org/education/saber. Countries can use these tools, which are based on the framework and indicator areas shown in Table 2, to gain a better understanding of their current status and needs in the area of student assessment and to plan for where to go next.

Stages of, and Strategies for, Development

The basic structure of the aforementioned rubrics for evaluating the data collected using the standardized questionnaires is summarized in Table 3. The goal of the rubrics is to provide a country with some sense of the development level of its assessment activities compared to best or recommended practice in the area.

Table 3. Basic Structure of Rubrics for Evaluating Data on Student Assessment Systems

	Development Level				
Theme/Dimension	LATENT (Absence of, or deviation from, attribute)	EMERGING (On way to meeting acceptable minimum standard)	ESTABLISHED (Acceptable minimum standard)	CUTTING EDGE (Best practice)	Rationale/Justification
EE – ENABLING CONTEXT					
EE1 – Policies					
EE2 – Fiscal resources					
EE3 – Organizational structures					
EE4 – Human resources					
SA – SYSTEM ALIGNMENT					
SA1 – Learning goals and standards					
SA2 – Curriculum					
SA3 – Pre-, in-service teacher training					
TQ – ASSESSMENT QUALITY					
TQ1 – Design					
TQ2 – Administration					

TQ3 – Analysis					
TQ4 – Uses					

Source: Author and M. Ramirez.

For each indicator, the rubric displays four development levels—*Latent*, *Emerging*, *Established,* and *Cutting Edge*.[20] Each level is accompanied by a description of what performance on the indicator looks like at that level. *Latent* is the lowest level of performance; it represents absence of, or deviation from, the attribute. *Emerging* is the next level; it represents partial presence of the attribute. *Established* represents the acceptable minimum standard on the indicator, and *Cutting Edge* represents the ideal or current best practice. Not all questions from the questionnaire are represented in the rubrics; this is because not all of the questions are underpinned by an evidence base that demonstrates a relationship between increasing performance levels on the attribute/indicator and improved quality of assessment activities.

In addition to evaluating performance on individual indicators, it can be useful to compare an assessment system's overall performance on the indicators against stylized vignettes or profiles of assessment systems as they look at different stages of development. Table 4 outlines some generic profiles for assessment systems at the *Emerging*, *Established*, and *Cutting Edge* stages of development (*Latent* is omitted because it basically represents the absence of any assessment activity).

[20] The *Latent* label could be applied to countries where there is no formal assessment activity or where the education system has been suspended due to war or other conflict.

Table 4. Stages of Student Assessment System Development

	Emerging	Established	Cutting Edge
Enabling context	• No or limited policy framework • Few trained staff; high turnover • Unreliable funding • Unclear or unstable institutional structures/arrangements	• Presence of policy framework • Training programs/trained staff with low turnover • Stable/reliable funding • Clear and stable institutional structures/arrangements	The same as for Established + strong focus on: ❖ Assessment for learning ❖ School-based and classroom assessment ❖ Role of teachers ❖ Innovation and research-based practices
System alignment	• Assessments not fully aligned with learning goals, standards, curriculum • Assessments not aligned with pre- and in-service teacher training opportunities • Limited use of results to inform policy and practice	• Assessments aligned with learning goals, standards, curriculum • Assessments aligned with pre- and in-service teacher training opportunities • Systematic use of results to inform policy and practice	
Assessment quality	• Limited awareness or application of technical or professional standards	• Awareness and application of technical or professional standards	

Source: **Author.**

Assessment systems that are at the *Emerging* stage tend to have enabling contexts, as well as levels of system alignment and assessment quality, that are just taking shape. These systems are characterized by instability and uncertainty about the choice, frequency, and use of assessment activities, indicative of an unclear vision for assessment at the system level and uncertain or insufficient funding for assessment activities. In this context, assessment is more

likely to function as an 'add on' to the system, without much systematic effort to align it with standards, curricula, or teacher training opportunities. Capacity building tends to be nonsystematic and of limited effectiveness as individuals disperse to other parts of the organization or to the private sector after they have been trained. Assessment activities tend to be of low quality due to a lack of awareness of, or attention to, professional standards.

Assessment systems that are at the *Established* stage tend to have enabling contexts, as well as levels of system alignment and assessment quality, that are stable, assured, or consolidated in nature. These systems are characterized by continuity and certainty about the choice, frequency, and use of assessment activities, as well as stable and sufficient sources of funding, indicative of a vision and 'buy in' for assessment at the system level. In this environment, assessment functions more as an integral part of the system, with systematic efforts to align it with standards, curricula, or teacher training opportunities. Capacity building tends to be focused, sustained, and effective and there is low staff turnover. Assessment activities tend to be of good quality due to awareness of, and attention to, professional standards. This stage may be viewed as the acceptable minimum standard in order for an assessment system to be truly effective.

Assessment systems that are at the *Cutting Edge* stage tend to have enabling contexts, as well as levels of system alignment and assessment quality that are highly developed in nature. In addition to having the features of *Established* systems, *Cutting Edge* systems are characterized by high levels of innovation and research-based practices as well as regular review of assessment activities. In this environment, assessment functions as a highly integral part of the

system. Capacity building tends to be very much focused on teachers, in addition to 'technicians,' testimony to a strong emphasis on school-based and classroom assessment.

It is worth noting that a system may be at different stages of development in relation to different types of assessment activity; that is, a system may be *Established* in the area of examinations but *Emerging* in the area of large-scale assessment, and vice versa. While it is generally better to be further along in as many areas as possible, it is not necessarily vital to be functioning at *Cutting Edge* levels in every aspect. Therefore, one might view the *Established* level as a desirable minimum outcome to achieve in all areas, but only aspire beyond that in those assessment areas that most contribute to the national vision or priorities for education.

While it is useful to have some sense of what assessment systems look like at different stages, it is just as important to understand how to progress *through* those stages. Thus, we also need to understand some of the key reforms or inputs that countries have used to develop more effective assessment systems.

The main factor that characterizes systems that make the shift from *Emerging* to *Established* is a concerted focus on reforms, inputs, and practices that strengthen the enabling context for assessment (Ferrer, 2006).[21] In their review of World Bank support for assessment projects in client countries, Larach and Lockheed (1992) found that projects that focused on improving institutional quality before addressing either assessment quality or dissemination issues were more likely to succeed than projects that first tried to improve assessment quality or

[21] While it may benefit a system, for a short time, to focus resources around making progress on one specific quality driver (for example, enabling context), this is not a long-term strategy as each quality driver is a necessary contributor to an effective assessment system.

dissemination. Similarly, in their review of assessment reform efforts in Central and Eastern European countries, West and Crighton (1999) noted that reforms had a better chance when there was public consensus that change was needed, clear and consistent political support for change, and sufficient allocation of resources.

The main factor that characterizes systems that make the shift from *Established* to *Cutting Edge* is a focus on reforms, inputs, and practices that prioritize the classroom, and teachers and students as the key actors in assessment (Darling-Hammond and Wentworth, 2010; Shepard, 2000).

Conclusions

This paper has extracted principles and guidelines from countries' experiences and the current research base to outline a framework for developing a more effective student assessment system. The framework provides policy makers and others with a structure for discussion and consensus building around priorities and key inputs for their assessment system.

While the value of this set of guidelines and principles should not be downplayed, it is important to also emphasize the significance of a county's own context, aspirations, and needs in deciding where to start, what approach to use, and how long to take. Countries should, therefore, view this framework as one that allows them a high degree of flexibility in what, when, and how they move forward. The measure of success at the end of the day is an assessment system that contributes to higher levels of education quality and student learning.

References

Airasian, P., and M. Russell. 2007. *Classroom Assessment: Concepts and Applications* (6th ed.). New York: McGrath Hill.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing.* Washington, DC: AERA.

Bishop, J., F. Mane, and M. Bishop. 2001. "Secondary Education in the United States: What Can Others Learn from Our Mistakes?" CAHRS Working Paper Series. Cornell Center for Advanced Human Resource Studies (CAHRS).

Black, P., and D. Wiliam. 2005. (1998). "Assessment and Classroom Learning." *Assessment in Education: Principles, Policy and Practice* 5(1): 7-73.

Braun, H., and A. Kanjee. 2006. "Using Assessment to Improve Education in Developing Nations." In J. Cohen, D. Bloom, and M. Malin, eds., *Educating All Children: A Global Agenda.* Cambridge, MA: American Academy of Arts and Sciences.

Bray, M., and L. Steward, eds. 1998. *Examination Systems in Small States: Comparative Perspectives on Policies, Models and Operations*. London: The Commonwealth Secretariat.

Brinkley, M., J. Guthrie, and T. Wyatt. 1991. *A Survey of National Assessment and Examination Practices in OECD Countries*. Lugano, Switzerland: OECD.

Carnoy, M., and S. Loeb. 2002. "Does External Accountability Affect Student Outcomes? A Cross-State Analysis." *Educational Evaluation and Policy Analysis* 24(4): 305-331.

Clarke, M. 2007. "State Responses to the No Child Left Behind Act: The Uncertain Link between Implementation and 'Proficiency for All'." In C. Kaestle and A. Lodewick, eds., *To Educate a Nation: Federal and National Strategies of School Reform* (pp. 144-174). University of Kansas Press.

Darling-Hammond, L., and L. Wentworth. 2010. *Benchmarking Learning Systems: Student Performance Assessment in International Context.* Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Ferrer, G. 2006. *Educational Assessment Systems in Latin America: Current Practice and Future Challenges.* Washington, DC: Partnership for Educational Revitalization in the Americas.

Fuhrman, S., and D. Elmore, eds. 1994. *Governing Curriculum*. Alexandria, VA: ASCD.

Greaney, V., and T. Kellaghan. 2008. *Assessing National Achievement Levels in Education.* Washington, DC: World Bank.

———. (1995). *Equity Issues in Public Examinations in Developing Countries.* Washington, DC: World Bank.

Hamilton, L., B. Stecher, and S. Klein., eds. 2002. *Making Sense of Test-Based Accountability in Education*. Santa Monica, CA: RAND Corporation.

Hanushek, E., and L. Woessmann. 2009. "Schooling, Cognitive Skills, and the Latin American Growth Puzzle." Working Paper 15066. Cambridge, MA: National Bureau of Economic Research.

———. 2007. *Education Quality and Economic Growth.* Washington, DC: World Bank.

Heubert, J., and R. Hauser. 1999. *High Stakes: Testing for Tracking, Promotion, and Graduation*. Washington, DC: National Academy Press.

Hoxby, C. 2002. "The Cost of Accountability." NBER Working Paper Series No. w8855. Cambridge, MA: National Bureau of Economic Research. Available at SSRN: http://ssrn.com/abstract=305599.

Independent Evaluation Group. 2006. *From Schooling Access to Learning Outcomes: An Unfinished Agenda.* Washington, DC: World Bank.

Kifer, E. 2001. *Large-Scale Assessment: Dimensions, Dilemmas, and Policy*. Thousand Oaks, CA: Corwin Press, Inc.

Larach, L., and M. Lockheed. 1992. "World Bank Lending for Educational Testing." PHREE Background Paper, 92/62R. Population and Human Resources Department. Washington, DC: World Bank.

Liberman, J. and M. Clarke. 2011. *Review of World Bank Support for Assessment Activities in Client Countries.* (draft manuscript). Washington, DC: World Bank.

Lockheed, M. 2009. *Review of Donor Support for Assessment Capacity Building in Developing Countries.* Unpublished manuscript. Washington, DC: World Bank.

Macintosh, H. 1994. *A Comparative Study of Current Theories and Practices in Assessing Students' Achievements at Primary and Secondary Level.* IBE Document Series, Number 4. Geneva, Switzerland: International Bureau of Education.

Madaus, G., and M. Clarke. 2001. "The Impact of High-Stakes Testing on Minority Students." In M. Kornhaber and G. Orfield, eds., *Raising Standards or Raising Barriers: Inequality and High Stakes Testing in Public Education* (pp. 85-106). New York: Century Foundation.

Madaus G., M. Clarke, and M. O'Leary. 2003. "A Century of Standardized Mathematics Testing." In G. M.A. Stanic and J. Kilpatrick, eds., *A History of School Mathematics* (pp. 1311-1434). Reston, VA: NCTM.

McKinsey & Company. 2007. *How the World's Best Performing School Systems Come Out On Top*. London: McKinsey & Company.

Messick, S. 1989. "Validity." In R. Linn, ed., *Educational Measurement* (3rd ed.) (pp. 13-103). New York: American Council on Education/Macmillan.

Organisation for Economic Co-operation and Development (OECD). 2010. *The High Cost of Low Educational Performance. The Long-Run Economic Impact of Improving PISA Outcomes.* Paris: OECD.

Ravela, P. 2005. "A Formative Approach to National Assessments: The Case of Uruguay." *Prospects* 35(1): 21-43.

Ravela, P., P. Arregui, G. Valverde, R. Wolfe, G. Ferrer, F. Martinez, M. Aylwin, and L. Wolff. 2008. "The Educational Assessments that Latin America Needs." Working Paper Series No. 40. Washington, DC: Partnership for Educational Revitalization in the Americas (PREAL).

Ravela, P., P. Arregui, G. Valverde, R. Wolfe, G. Ferrer, F. M. Rizo, M. Aylwin, and L. Wolff. 2009. "The Educational Assessments that Latin America Needs." Washington, DC: PREAL.

Rodriguez, M. C. 2004. "The Role of Classroom Assessment in Student Performance on TIMSS." *Applied Measurement in Education* 17(1): 1-24.

Shepard, L. 2000. "The Role of Assessment in a Learning Culture." *Educational Researcher* 29(7): 4-14.

Smith, M. S., and J. O'Day. 1991. "Systemic School Reform." In S.H. Fuhrman and B. Malen, eds., *The Politics of Curriculum and Testing, 1990 Yearbook of the Politics of Education Association* (pp. 233-267). London and Washington, DC: Falmer Press.

United Nations Educational, Scientific and Cultural Organization (UNESCO). 2007. *Education for All Global Monitoring Report 2008: Education for All by 2015. Will We Make It?* Paris: UNESCO/Oxford University Press.

West, R., and J. Crighton. 1999. "Examination Reform in Central and Eastern Europe: Issues and Trends." *Assessment in Education* 6(2): 271-280.

Wolff, L. 2007. *The Costs of Student Assessment in Latin America*. Washington, DC: PREAL.

World Bank. 2010. *Russia Education Aid for Development (READ) Trust Fund Annual Report 2009.* Washington, DC: World Bank.

Annex 1. Assessment Types and Their Key Differences

	Classroom	Large-scale assessment surveys		Examinations	
		National	International	Exit	Entrance
Purpose	• To provide immediate feedback to inform classroom instruction	• To provide feedback on overall health of the system at particular grade/age level(s), and to monitor trends in learning	• To provide feedback on the comparative performance of the education system at particular grade/age level(s)	• To certify students as they move from one level of the education system to the next (or into the workforce)	• To select students for further educational opportunities
Frequency	• Daily	• For individual subjects offered on a regular basis (such as every 3-5 years)	• For individual subjects offered on a regular basis (such as every 3-5 years)	• Annually and more often where the system allows for repeats	• Annually and more often where the system allows for repeats
Who is tested?	• All students	• Sample or census of students at a particular grade or age level(s)	• A sample of students at a particular grade or age level(s)	• All eligible students	• All eligible students
Format	• Varies from observation to questioning to paper-and-pencil tests to student performances	• Usually multiple choice and short answer	• Usually multiple choice and short answer	• Usually essay and multiple choice	• Usually essay and multiple choice
Coverage of curriculum	• All subject areas	• Generally confined to a few subjects	• Generally confined to one or two subjects	• Covers main subject areas	• Covers main subject areas
Additional information collected from students?	• Yes, as part of the teaching process	• Frequently	• Yes	• Seldom	• Seldom
Scoring	• Usually informal and simple	• Varies from simple to more statistically sophisticated techniques	• Usually involves statistically sophisticated techniques	• Varies from simple to more statistically sophisticated techniques	• Varies from simple to more statistically sophisticated techniques

Source: Author.