SYSTEMS APPROACH FOR BETTER EDUCATION RESULTS

SABER
Working Paper Series

Number 1 April 2012

What Matters Most for Student Assessment Systems: A Framework Paper

Marguerite Clarke

THE WORLD BANK

Contents

About the Series

Building strong education systems that promote learning is fundamental to development and economic growth. Over the past few years, as developing countries have succeeded in building more classrooms, and getting millions more children into school, the education community has begun to actively embrace the vision of measurable learning for all children in school. However, learning depends not only on resources invested in the school system, but also on the quality of the policies and institutions that enable their use and on how well the policies are implemented.

In 2011, the World Bank Group launched Education Sector Strategy 2020: Learning for All, which outlines an agenda for achieving "Learning for All" in the developing world over the next decade. To support implementation of the strategy, the World Bank commenced a multi-year program to support countries in systematically examining and strengthening the performance of their education systems. This evidence-based initiative, called SABER (Systems Approach for Better Education Results), is building a toolkit of diagnostics for examining education systems and their component policy domains against global standards, best practices, and in comparison with the policies and practices of countries around the world. By leveraging this global knowledge, SABER fills a gap in the availability of data and evidence on what matters most to improve the quality of education and achievement of better results.

SABER-Student Assessment, one of the systems examined within the SABER program, has developed tools to analyze and benchmark student assessment policies and systems around the world, with the goal of promoting stronger assessment systems that contribute to improved education quality and learning for all. To help explore the state of knowledge in the area, the SABER-Student Assessment team invited leading academics, assessment experts, and practitioners from developing and industrialized countries to come together to discuss assessment issues relevant for improving education quality and learning outcomes. The papers and case studies on student assessment in this series are the result of those conversations and the underlying research. Prior to publication, all of the papers benefited from a rigorous review process, which included comments from World Bank staff, academics, development practitioners, and country assessment experts.

All SABER-Student Assessment papers in this series were made possible by support from the Russia Education Aid for Development Trust Fund (READ TF). READ TF is a collaboration between the Russian Federation and the World Bank that supports the improvement of student learning outcomes in low-income countries through the development of robust student assessment systems.

The SABER working paper series was produced under the general guidance of Elizabeth King, Education Director, and Robin Horn, Education Manager in the Human Development Network of the World Bank. The Student Assessment papers in the series were produced under the technical leadership of Marguerite Clarke, Senior Education Specialist and SABER-Student Assessment Team Coordinator in the Human Development Network of the World Bank. Papers in this series represent the independent views of the authors.

About the Author

Marguerite Clarke is a Senior Education Specialist in the Human Development Network at the World Bank. She leads the Bank's work on learning assessment, including providing support to individual countries to improve their assessment activities and uses of assessment information, and heading the global work program on student assessment under the Russia Education Aid for Development (READ) Trust Fund program. Under READ, she is responsible for developing evidence-based tools and approaches for evaluating and strengthening the quality of student assessment systems. Prior to joining the Bank, Marguerite was involved in research, policy, and practice in the areas of higher education teaching and learning, higher education quality, and student assessment and testing policy at universities in Australia (University of South Australia) and the United States (Brown University, Boston College). She also worked as a classroom teacher in the Chinese, Irish, Japanese, and U.S. education systems and received a national teaching award from the Irish Department of Education in 1989. A former Fulbright Scholar, she received her PhD in Educational Research, Measurement, and Evaluation from Boston College (2000) and is on the advisory board of the UNESCO Institute for Statistics Observatory for Learning Outcomes.

Acknowledgments

Many people provided inputs and suggestions for this paper. Thanks in particular go to the peer reviewers and meeting chairs: Luis Benveniste, Luis Crouch, Deon Filmer, Robin Horn, Elizabeth King, Marlaine Lockheed, Harry Patrinos, and Alberto Rodriguez. I am also grateful to the READ Trust Fund team, particularly Julia Liberman and María-José Ramírez, who provided valuable support in developing a set of rubrics and questionnaires based on this framework paper, as well as Olav Christensen, Emily Gardner, Manorama Gotur, Emine Kildirgici, Diana Manevskaya, Cassia Miranda, and Fahma Nur. Thanks also to READ Technical Group members, past and present, including Luis Benveniste, Cedric Croft, Amber Gove, Vincent Greaney, Anil Kanjee, Thomas Kellaghan, Marina Kuznetsova, María-José Ramírez, and Yulia Tumeneva, as well as to the Task Team Leaders and teams in the READ countries. Others who provided helpful insights and suggestions along the way include Patricia Arregui, Felipe Barrera, Viktor Bolotov, Lester Flockton, Alejandro Ganimian, Juliana Guaqueta, Gabrielle Matters, Emilio Porta, Halsey Rogers, Alan Ruby, Jee-Peng Tan, Igor Valdman, and Emiliana Vegas. Special thanks to the Russian government for their support for this work under the READ Trust Fund program.

Abstract

The purpose of this paper is to provide an overview of *what matters most for building a more effective student assessment system*. The focus is on systems for assessing student learning and achievement at the primary and secondary levels.[1] The paper extracts principles and guidelines from countries' experiences, professional testing standards, and the current research base. The goal is to provide national policy makers, education ministry officials, development organization staff, and other stakeholders with a *framework and key indicators for diagnosis, discussion, and consensus-building around how to construct a sound and sustainable student assessment system that will support improved education quality and learning for all*.

[1] This paper does not discuss psychological or workplace testing; nor does it explicitly discuss assessment of student learning and achievement at the tertiary level, although many of the issues raised also apply to that level of schooling.

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. 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 based on the results may vary from how to design system-wide programs to improve teaching and learning in schools, 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 and achievement. 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 education quality and student learning outcomes (Ravela et al., 2009).[4] Meeting these information and decision-making needs in a way that has the support of key political and other groups in society will contribute to the longer-term sustainability and effectiveness of the assessment system.

Governments, international organizations, and other stakeholders are increasingly recognizing the importance of assessment for monitoring and improving student learning and achievement levels, and the concomitant need to develop strong systems for student assessment (IEG, 2006; McKinsey & Company, 2007; UNESCO, 2007). This recognition is linked to growing evidence that many of the benefits of education—cultural, economic, and social—accrue to society only when learning occurs (OECD, 2010). For example, an increase of one standard deviation in scores on international assessments of reading and mathematics achievement levels has been linked to a 2 percent increase in annual growth rates of GDP per capita (Hanushek and Woessmann, 2007, 2009).

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*, typically 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. While these cost efficiencies are appealing, they should not be allowed to obscure other important factors—for example, equity and social goals—that need to be considered in any decision about whether or not to implement a particular assessment program.

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

---

[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 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 education system.

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).[5] Nongovernmental organizations also have increasingly turned to student assessment to draw public attention to poor achievement levels and to create an impetus for change.

Despite this interest in student assessment, 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 and achievement levels, 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 only provides a snapshot of student achievement levels while the latter allows for the possibility of monitoring trends in achievement and learning levels 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 have shock value and create an opening for discussions about education quality, and this can be a short-term strategy for putting learning on the agenda.[6] 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 and achievement. This is the only way to harness the full power of assessment.

Theory and Evidence on Student Assessment

A basic premise of the research on student assessment is that the right kinds of assessment activities, and the right uses of the data generated by those activities, contribute to better outcomes, be those improved learning or improved policy decisions (for example, Heubert and Hauser, 1999).[7] What constitutes 'right' is largely driven by a set of theoretical and technical guidelines for test developers and users of assessment information (AERA, APA, and NCME, 1999).

---

[5] 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 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.

[6] 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."

[7] Ravela et al. (2008) note that student assessment is a necessary, but insufficient, condition for improving education. There is some evidence that the mere existence and dissemination of assessment 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.

There also is a sizeable body of empirical research showing the benefits of specific types of assessment activities, when implemented and used correctly, on student learning. For example, research demonstrates a strong link between high-quality, formative classroom assessment activities and better student learning outcomes as measured by student performance on standardized tests of educational achievement. Black and Wiliam's (1998) synthesis of over 250 empirical studies from around the world on the impact of high-quality, formative classroom assessment activities shows student gains of a half to a full standard deviation on standardized achievement tests, with the largest gains being realized by low achievers.[8] Black and Wiliam (1998) conclude:

> The gains in achievement appear to be quite considerable, and … amongst the largest ever reported for educational interventions. As an illustration of just how big these gains are, an effect size of 0.7, if it could be achieved on a nationwide scale, would be equivalent to raising the mathematics attainment score of an "average" country like England, New Zealand or the United States into the "top five" after the Pacific rim countries of Singapore, Korea, Japan and Hong Kong. (p. 61)

Bennett (2011), however, notes that more work needs to be done to define and isolate the specific characteristics of formative classroom assessment activities that lead to improved student learning outcomes.[9]

Correlational research on high school or upper-secondary exit examinations demonstrates a link between countries that have those policies and higher student performance levels on international assessments, such as PISA or TIMSS (for example, Bishop, Mane and Bishop, 2001). Other studies show a link between specific characteristics of the tests used in these examination programs and student learning outcomes, with curriculum- or subject-based examinations (as opposed to more general ability or aptitude tests) viewed as most effective in promoting better student learning outcomes (Au, 2007; Hill, 2010).

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

Results from large-scale, system-level assessments of overall student achievement levels increasingly provide the foundation for test-based accountability programs in many countries. Research shows an overall weak, but positive, link between the uses of data from these assessments to hold schools and educators accountable (through, for example, league tables, monetary rewards, or staffing decisions) and better student learning outcomes (for example, Carnoy and Loeb, 2002). At the same time, findings

---

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

[9] One meta-analysis of 21 controlled studies (Fuchs and Fuchs, 1986) that looked at the *frequency* of classroom assessment activities found that systematic use of formative classroom assessment activities—weekly or even more often—can have a strong positive effect on student achievement (for example, two assessments per week results in an effect size of 0.85, or a percentile gain of 30 points).

suggest that simply reporting information about average school scores on these assessments also can lead to increased student performance (Hanushek and Raymond, 2003), suggesting that there still is much to learn about the optimal mix of incentives for test-based accountability models that will produce the best outcomes with the fewest negative side effects. To date, research suggests that key determinants of whether the effects of test-based accountability exercises are more positive than negative include the technical quality of the tests themselves, the alignment between the test design and the way test results are used, and the extent to which supports are in place to help schools or teachers identified as underperforming (Ravela, 2005).[10]

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, system-level assessments). This research draws on *principles and best practices* 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 education systems around the world (for example, Australia, Finland, Singapore, Sweden, and the United Kingdom) and noted that student assessment activities in these systems:

- illustrate the importance of assessment *of*, *for,* and *as* student 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[11]
- 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.[12]

---

[10] 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.

[11] That is to say, some countries have good outcomes on international assessment exercises, but don't use a lot of standardized testing in their own education systems (for example, Finland). Other countries place a lot of emphasis on standardized testing (for example, the United States), but don't do so well on the same international assessment exercises.

[12] Results from standardized performance tasks are incorporated into students' 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

While Darling-Hammond and Wentworth's research provides a broad vision of what an effective assessment system looks like, it does not tell us what it takes to get there. Other studies delve into these planning, process, and implementation 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 analysis for secondary school examinations. Others (for example, Lockheed, 2009) evaluate the status of donor activity in the area of assessment and discuss how to improve the effectiveness of this support to countries. Still others delve into the politics of creating sustainable and effective assessment systems (McDermott, 2011).

This paper draws together all of the above streams of evidence, organizing the key issues and factors into 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 the framework in a strategic way, we need to identify some key dimensions of assessment systems. Two main dimensions are discussed here: (i) *types/purposes* of assessment activities and (ii) the *quality* of those activities.

Dimension 1. Assessment Types/Purposes

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

- *classroom assessments* for providing real-time information to support teaching and learning in individual classrooms
- *examinations* for making decisions about an individual student's progress through the education system (for example, certification or selection decisions), including the allocation of 'scarce' educational opportunities
- *large-scale, system-level assessments* for monitoring and providing policy-maker- 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 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 or formative 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 nonstandardized instruments and procedures for collecting and interpreting written, oral, and other forms of evidence on student learning or achievement. Examples of classroom assessment activities include oral questioning and feedback, homework assignments, student

---

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.

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.

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 or exit examinations at the end of compulsory education in many education systems are a good example. As discussed earlier, the high-stakes nature of most examinations means they can exert a backwash effect on the education system in terms of what is taught and learned, having an impact, for better or worse, on the skills and knowledge profile of graduates (West and Crighton, 1999). Such consequences must be considered when determining whether the use of such tests is appropriate[13] and whether or how they should be combined with other sources of information in order to ensure that the results are used in a way that is as fair as possible to individuals, groups, and society as a whole. 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).

Large-scale, system-level 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 education 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 the state-level tests in the United States or Canada.[14] These assessments vary in the grades or age levels tested, coverage of the target population (sample or census), internal or external focus (for example, national versus international benchmarks), 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, as discussed earlier, while some stop at the

[13] 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.

[14] 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.

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).[15]

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 the Early Grade Reading Assessment (EGRA), 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 (Gove and Cvelich, 2011).

Education systems can have quite different profiles in terms of the emphasis placed on the different types of assessment activities. For example, Finland's education system emphasizes classroom assessment as the key source of information on student learning and achievement and draws far less on examinations or large-scale, system-level 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 surveys (although this is changing).[16] Factors contributing to these different assessment system profiles vary from the official vision and goals of the education system (and the role of assessment in achieving that vision) to the economic structures and opportunities in a country and the related information needs of key stakeholders. It is not clear that there exists *one* ideal profile for an assessment system that works equally well in all contexts.

Dimension 2. Quality Drivers

Instead of being able to reference one ideal profile for a student 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

[15] World Bank support for assessment activity over the last 20 years (Larch and Lockheed, 1992; Liberman and Clarke, 2012) has shifted from an emphasis on examination reform to an emphasis on the implementation of large-scale, system-level assessment exercises for monitoring achievement trends and informing policy and practice.

[16] Other contributing factors include 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); the capacity of various stakeholders in the system to effectively carry out different types of assessment activities (Greaney and Kellaghan, 2008); and the cost, perceived or real, of assessment activities (Wolff, 2007).

- assessment quality.

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

The *enabling context* refers to the broader context in which an 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 legislative or policy framework for assessment activities; leadership surrounding the assessment activity (including the political will to implement an assessment in spite of the knowledge that results might reveal serious issues or inequities in student learning); public engagement with the assessment activity; the institutional arrangements for designing, carrying out, or using the results from the assessment activity;[17] 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 (at the same time, keeping in mind that relative autonomy from political influence is one of the hallmarks of a more mature assessment system), which may subsequently be implemented via public-private partnerships (for example, contracting administration of an assessment program to an outside firm). 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 (for example, when states or systems come together to design a common test item bank that each can use for their own purposes, hence reducing the cost for individual states or systems). 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 is aligned or coherent with other components 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 provide is of use to improving the quality of education in the system, and so that synergies can be created.

Alignment involves more than a simple match between what is tested and what is in the official standards or intended curriculum (at the same time, it is important that most assessment activities provide at least some information on student learning and achievement in relation to official standards or 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 and aspirations for the education system and related

[17] There is much debate over whether examination or large-scale 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 the 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.

reforms. Under such a scenario, assessment can actually lead quality improvements in the education system rather than simply passively monitor them (notwithstanding that the use of data from TIMSS, PIRLS, and PISA to monitor the impact of national 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, system-level 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). In fact, poor assessment quality could undermine the entire assessment exercise if it causes distrust in the approach.

Two technical issues that need to be considered in any review of assessment quality 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 diagramed in table 1.

Table 1. Framework for Building a More Effective Student Assessment System

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

Source: World Bank.

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 begin a discussion about where the system appears strong and where more work may be 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
- 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
- theory—that is, general consensus among experts that it contributes to effective assessment.

The evidence base is stronger in some areas than in others. For example, there are many professional standards for assessment quality (APA, AERA, and NCME, 1999),[18] but far fewer for the enabling context. In addition, some of the empirical research is limited by its correlational nature and hence we must be cautious about inappropriate attribution or over-interpreting the association between characteristics. Despite such limitations, evidence from a variety of sources converges quite convincingly to make clear what better assessment is (and what it is not).

The above criteria and considerations were used to expand the three quality drivers into the broad indicator areas shown in table 2. These indicator areas are most relevant to examinations and large-scale, system-level assessment activities, but also can be applied to classroom assessment.

Table 2. Framework for Building a More Effective Student Assessment System, with Broad Indicator Areas

	Assessment types/purposes		
	Classroom assessment	**Examinations**	**Large-scale, system-level assessment**
Enabling context	Policies Leadership and public engagement Funding Institutional arrangements Human resources		
System alignment	Learning/quality goals Curriculum Pre- and in-service teacher training opportunities		
Assessment quality	Ensuring quality (design, administration, analysis) Ensuring effective uses		

Source: World Bank.

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

Data pertaining to some of these indicator areas can be found in official documents, published reports (for example, Ferrer, 2006), research articles (for example, Braun and Kanjee, 2006), and online databases.[19] For the most part, however, the relevant data have not been gathered in any comprehensive or systematic fashion.[20] Those wishing to review this type of information for a particular assessment system most likely will need to collect the data themselves. In response to this need, the World Bank has developed a set of standardized questionnaires and rubrics for collecting and evaluating data on the three assessment types (classroom assessments, examinations, and large-scale, system-level assessment) and related quality drivers (enabling context, system alignment, assessment quality). The tools, which are regularly updated on the basis of new evidence and country experiences, are available at http://www.worldbank.org/education/saber. Countries can use these tools, which build on the framework and broad indicator areas shown in table 2, to systematically examine and gain a better understanding of the strengths and weaknesses of their student assessment system and to plan for where to go next. It is important to point out that *the tools primarily focus on benchmarking a country's policies, practices, and arrangements* for classroom assessment, examinations, and large-scale, system-level assessment activities *at the system-level*. Additional tools would be needed to determine actual, on-the-ground practices by teachers and students in schools.

Levels of Development

The basic structure of the rubrics for evaluating data collected using the standardized questionnaires is summarized in table 3. The full set of rubrics is provided in appendix 2. 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.

[19] Two of the more useful online databases are http://www.inca.org.uk/ and http://epdc.org/.

[20] Brinkley, Guthrie, and Wyatt (1991) surveyed large-scale, system-level 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).

Table 3. Basic Structure of Rubrics for Evaluating Data Collected on a Student Assessment System

Dimension	Development Level				
	LATENT (Absence of, or deviation from, attribute)	**EMERGING (On way to meeting minimum standard)**	**ESTABLISHED (Acceptable minimum standard)**	**ADVANCED (Best practice)**	**Justification**
EC—ENABLING CONTEXT					
EC1—Policies					
EC2—Leadership, public engagement					
EC3—Funding					
EC4—Institutional arrangements					
EC5—Human resources					
SA—SYSTEM ALIGNMENT					
SA1—Learning/quality goals					
SA2—Curriculum					
SA3—Pre-, in-service teacher training					
AQ—ASSESSMENT QUALITY					
AQ1—Ensuring quality (design, administration, analysis)					
AQ2—Ensuring effective uses					

Source: World Bank.

For each indicator, the rubric displays four development levels—*Latent*, *Emerging*, *Established,* and *Advanced*.[21] These levels are artificially constructed categories chosen to represent key stages on the underlying continuum for each indicator. 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 *Advanced* represents the ideal or current best practice. Not all questions from the questionnaires are represented in the rubrics; this is because not all of the questions are underpinned by an evidence base that

[21] 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.

demonstrates a relationship between increasing performance levels on the attribute/indicator and improved quality or effectiveness of assessment activities.

It is important to recognize that many of the issues that we are trying to get at with the indicators and associated development levels can be difficult to measure. In some instances, explicit technical standards exist and can be drawn on to aid these measurement efforts (for example, international standards for determining whether a country's TIMSS results are sufficiently robust to be included in the international report). In others, judgment calls need to be made (for example, measuring the degree of public support for a particular assessment activity). In order to enhance the overall reliability and cross-system comparability of the indicators and development levels, the questionnaires and rubrics rely, as much as possible, on objective measures.

In addition to evaluating performance on individual indicators, it can be useful to qualitatively compare an assessment system's overall characteristics against profiles of assessment systems as they might look at different levels of development. Table 4 outlines generic profiles—drawing on the information provided in table 2 and appendix 2—for assessment systems at *Emerging*, *Established,* and *Advanced* levels of development (*Latent* is omitted because it basically represents the absence of any assessment activity).

Assessment systems that are at an *Emerging* level can be characterized as having 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.

Table 4. Stylized Profiles of Student Assessment Systems at Different Levels of Development

	Emerging	Established	Advanced
Enabling context	• No or limited policy framework or guidelines • Weak leadership/public engagement • Few trained staff; high turnover • Unreliable/irregular funding • Unclear or unstable institutional arrangements	• Presence of clear policy framework or guidelines • Strong leadership/public engagement • Training programs/trained staff with low turnover • Stable/regular funding • Clear and stable institutional 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/quality goals, standards, curriculum • Assessments not aligned with pre- and in-service teacher training opportunities	• Assessments aligned with learning/quality goals, standards, curriculum • Assessments aligned with pre- and in-service teacher training opportunities	
Assessment quality	• Limited awareness or application of technical or professional standards for ensuring assessment quality and effective uses	• Awareness and application of technical or professional standards for ensuring assessment quality effective uses	

Source: World Bank.
Note: The *Latent* level is omitted because it basically represents the absence of any assessment activity.

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 an *Established* level can be characterized as having 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 level may be viewed as the acceptable minimum standard in order for an assessment system to be effective.

Assessment systems that are at an *Advanced* level can be characterized as having enabling contexts, as well as levels of system alignment and assessment quality that are highly developed in nature. In addition to having the best features of *Established* systems, *Advanced* systems are characterized by high

levels of innovation and research-based practices. 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 (and reminiscent of the key features of high-performing systems highlighted by Darling-Hammond and Wentworth in their work).

In reality, assessment systems are likely to be at different levels of development in different areas. For example, a system may be *Established* in the area of examinations, but *Emerging* in the area of large-scale, system-level assessment, and vice versa. While intuition suggests that it is probably better to be further along in as many areas as possible, the evidence is unclear as to whether it is necessary to be functioning at *Advanced* levels in all areas. Therefore, one might view the *Established* level as a desirable minimum outcome to achieve in all areas (which is what we see in the assessment systems of countries like Finland and Australia), but only aspire beyond that in those areas that most contribute to the national vision or priorities for education. In line with these considerations, the ratings generated by the rubrics in appendix 2 are not meant to be additive across assessment types (that is, they are not meant to be added to create an overall rating for an assessment system; they are only meant to produce an overall rating for each assessment type).

While it is useful to have an idea of what assessment systems and different assessment types look like at different development levels, it is equally, if not more, useful to know how to progress *through* those levels. Thus, we also need to understand some of the key reforms or inputs that countries have used to develop more effective assessment systems. Unfortunately, the evidence becomes sparser in this area and further research is definitely needed to flesh out the concrete strategies involved.

Based on the small amount of available evidence, the main factor that seems to characterize systems that make the shift from *Emerging* to *Established* (overall or in a specific assessment area) is a concerted focus on reforms, inputs, and practices that strengthen the enabling context for assessment (Ferrer, 2006).[22] For example, in their review of World Bank support for assessment projects in client countries, Larach and Lockheed (1992) found that projects that first focused on improving institutional arrangements were more likely to succeed—in terms of leading to a sustainable assessment program in the country—than projects that first tried to improve the technical quality of existing assessment activities. In line with this finding, in their review of assessment reform efforts in Central and Eastern European countries, West and Crighton (1999) noted that reforms had a better chance of being sustained 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 seems to characterize systems that make the shift from *Established* to *Advanced* 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). This relates to the fact that the most powerful form of assessment, when done correctly, is that carried out by teachers and students in the course of their daily classroom activities (that is, classroom assessment). Doing this type of assessment correctly requires a lot of capacity building and focused attention on teacher quality issues.

[22] 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.

Conclusion

Assessment is key to knowing whether an education system is producing the desired outcomes for students, the economy, and society at large. Without effective assessment, it is impossible to know whether students are learning and whether reforms are working in the intended ways.

This paper 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 an evidence-based structure for discussion and consensus building around priorities and key inputs for their assessment system.

An important contribution of the framework is to help countries identify the key quality drivers that need to be addressed in order to strengthen the quality and utility of the information produced by the various activities in their assessment system. This is critical because the main purpose of any assessment system is to provide valid and timely information to a set of users—the student, the teacher, the community, and the policy maker—so that they can make better decisions in support of improved quality and learning outcomes. Choices about the assessment system need to be consistent with serving these users and their information and decision-making needs.

The framework also has a dynamic dimension that illustrates the trajectory of moving from one level of development to the next in each assessment area. It is important to keep in mind that it takes time to progress from level to level. Case studies on countries' experiences in strengthening their student assessment systems reveal that it often takes a decade or more for a set of reforms and inputs to really take hold and produce tangible results. Therefore, country teams must plan from the outset to have a long-term commitment to, and investment in, the policies, inputs, and actions that will be required to transform their assessment system. The payoff will be an assessment system that can support better decision making and contribute to higher levels of education quality and learning for all.

References

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

Au, W. 2007. "High-Stakes Testing and Curricular Control: A Qualitative Metasynthesis." *Educational Researcher* 36(5): 258–67.

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

Bennett, R. E. 2011. "Formative Assessment: A Critical Review." *Assessment in Education: Principles, Policy and Practice* 18(1): 5–25.

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. 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). Lawrence: 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.

Fuchs, L. S., and D. Fuchs. 1986. "Effects of Systematic Formative Evaluation on Student Achievement: A Meta-Analysis." *Exceptional Children* 53: 199–208.

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

Gove, A., and P. Cvelich. 2011. *Early Reading: Igniting Education for All. A Report by the Early Grade Learning Community of Practice. Revised Edition.* Research Triangle Park, NC: Research Triangle Institute.

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 M. Raymond. 2003. "Lessons about the Design of State Accountability Systems." In P. Peterson and M. West, eds., *No Child Left Behind? The Politics and Practice of Accountability* (pp. 127–151). Washington, DC: Brookings Institution Press.

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.

Hill, P. 2010. *Examination Systems.* Asia-Pacific Secondary Education System Review Series. Bangkok: UNESCO.

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 (IEG). 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. 2012. *Review of World Bank Support for Assessment Activities in Client Countries.* Unpublished 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.

McDermott, K. A. 2011. *High-Stakes Reform: The Politics of Educational Accountability.* Washington, DC: Georgetown University Press.

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.

Appendix 1: Assessment Types and Their Key Differences

		Large-scale, system-level assessment		
	Classroom	**National**	**International**	**Examinations**
Purpose	To provide immediate feedback to inform classroom instruction	To provide feedback on the 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 select or certify students as they move from one level of the education system to the next (or into the workforce)
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
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
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
Coverage of curriculum	All subject areas	Generally confined to a few subjects	Generally confined to one or two subjects	Covers main subject areas
Additional information collected from students?	Yes, as part of the teaching process	Frequently	Yes	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

Source: World Bank.

Appendix 2. Rubrics for Judging the Development Level of Different Assessment Types

Classroom Assessment

LATENT Absence of, or deviation from, the attribute	EMERGING On way to meeting minimum standard	ESTABLISHED Acceptable minimum standard	ADVANCED Best practice	Justification
Enabling Context & System Alignment (EC & SA) Overall policy and resource framework within which classroom assessment activity takes place in an education system, and the degree to which classroom assessment activity is coherent with other components of the education system.				
EC&SA1—Setting clear guidelines for classroom assessment				
(Q1) There is no system-level document that provides guidelines for classroom assessment.	(Q1) There is an informal system-level document that provides guidelines for classroom assessment.	(Q1) There is a formal system-level document that provides guidelines for classroom assessment.	This option does not apply to this dimension.	
This option does not apply to this dimension.	This option does not apply to this dimension.	(Q3, Q4) The availability of the document is restricted.	(Q3, Q4) The document is widely available.	
EC&SA2—Aligning classroom assessment with system learning goals				
(Q5) There are no system-wide resources for teachers for classroom assessment.	(Q5) There are scarce system-wide resources for teachers for classroom assessment.	(Q5) There are some system-wide resources for teachers for classroom assessment.	(Q5) There are a variety of system-wide resources available for teachers for classroom assessment.	
(Q6) There is no official curriculum or standards document.	(Q6) There is an official curriculum or standards document, but it is not clear what students are expected to learn or to what level of performance.	(Q6) There is an official curriculum or standards document that specifies what students are expected to learn, but the level of performance required is not clear.	(Q6) There is an official curriculum or standards document that specifies what students are expected to learn and to what level of performance.	
EC&SA3—Having effective human resources to carry out classroom assessment activities				
(Q7, Q8) There are no system-level mechanisms to ensure that teachers develop skills and expertise in classroom assessment.	This option does not apply to this dimension.	(Q7, Q8) There are some system-level mechanisms to ensure that teachers develop skills and expertise in classroom assessment.	(Q7, Q8) There are a variety of system-level mechanisms to ensure that teachers develop skills and expertise in classroom assessment.	

LATENT Absence of, or deviation from, the attribute	EMERGING On way to meeting minimum standard	ESTABLISHED Acceptable minimum standard	ADVANCED Best practice	Justification
Assessment Quality (AQ) Quality of classroom assessment design, administration, analysis, and use				
AQ1—Ensuring the quality of classroom assessment				
(Q11) Classroom assessment practices suffer from widespread weaknesses, or there is no information available on classroom assessment practices.	(Q11) Classroom assessment practices are known to be weak.	(Q11) Classroom assessment practices are known to be of moderate quality.	(Q11) Classroom assessment practices are known to be generally of high quality.	
(Q12) There are no mechanisms to monitor the quality of classroom assessment practices.	(Q12) There are ad hoc mechanisms to monitor the quality of classroom assessment practices.	(Q12) There are limited systematic mechanisms to monitor the quality of classroom assessment practices.	(Q12) There are varied and systematic mechanisms in place to monitor the quality of classroom assessment practices.	
AQ2—Ensuring effective uses of classroom assessment				
(Q14) Classroom assessment information is not required to be disseminated to key stakeholders.	This option does not apply to this dimension.	(Q14) Classroom assessment information is required to be disseminated to some key stakeholders.	(Q14) Classroom assessment information is required to be disseminated to all key stakeholders.	
(Q15) There are no required uses of classroom assessment to support student learning.	(Q15) There are limited required uses of classroom assessment to support student learning.	(Q15) There are adequate required uses of classroom assessment to support student learning, excluding its use as an input for external examination results.	(Q15) There are adequate required uses of classroom assessment to support student learning, including its use as an input for external examination results.	

Source: World Bank.

Examinations

LATENT Absence of, or deviation from, the attribute	EMERGING On way to meeting minimum standard	ESTABLISHED Acceptable minimum standard	ADVANCED Best practice	Justification
Enabling Context (EC) Overall framework of policies, leadership, organizational structures, fiscal, and human resources in which assessment activity takes place in an education system and the extent to which that framework is conducive to, or supportive of, the assessment activity.				
EC1—Setting clear policies				
(Q3_III) No standardized examination has taken place.	(Q3_III) The standardized examination has been operating on an irregular basis.	(Q3_III) The examination is a stable program that has been operating regularly.	This option does not apply to this dimension.	
(Q3) There is no policy document that authorizes the examination.	(Q3) There is an informal or draft policy document that authorizes the examination.	(Q3) There is a formal policy document that authorizes the examination.	This option does not apply to this dimension.	
This option does not apply to this dimension.	(Q5) The policy document is not available to the public.	(Q5) The policy document is available to the public.	This option does not apply to this dimension.	
This option does not apply to this dimension.	This option does not apply to this dimension.	(Q6) The policy document addresses some key aspects of the examination.	(Q6) The policy document addresses all key aspects of the examination.	
EC2—Having strong leadership				
(Q8) All stakeholder groups strongly oppose the examination.	(Q8) Most stakeholder groups oppose the examination.	(Q8) Most stakeholders groups support the examination.	(Q8) All stakeholder groups support the examination.	
(Q9) There are no attempts to improve the examination by stakeholder groups.	This option does not apply to this dimension.	(Q9) There are independent attempts to improve the examination by stakeholder groups.	(Q9) There are coordinated attempts to improve the examination by stakeholder groups.	
(Q10) Efforts to improve the examination are not welcomed by the leadership in charge of the examination.	This option does not apply to this dimension.	(Q10) Efforts to improve the examination are generally welcomed by the leadership in charge of the examination.	This option does not apply to this dimension.	
EC3—Having regular funding				
(Q11) There is no funding allocated for the examination.	(Q11) There is irregular funding allocated for the examination.	(Q11) There is regular funding allocated for the examination.	This option does not apply to this dimension.	
This option does not apply to this dimension.	(Q12) Funding covers some core examination activities: design, administration, data processing or reporting.	(Q12) Funding covers all core examination activities: design, administration, data processing, and reporting.	This option does not apply to this dimension.	
This option does not apply to this dimension.	(Q12) Funding does not cover research and development.	Does not apply.	(Q12) Funding covers research and development.	

LATENT Absence of, or deviation from, the attribute	EMERGING On way to meeting minimum standard	ESTABLISHED Acceptable minimum standard	ADVANCED Best practice	Justification
EC4—Having strong organizational structures				
(Q14) The examination office does not exist or is newly established.	(Q14) The examination office is newly established.	(Q14) The examination office is a stable organization.	This option does not apply to this dimension.	
(Q15) The examination office is not accountable to an external board or agency.	This option does not apply to this dimension.	(Q15) The examination office is accountable to an external board or agency.	This option does not apply to this dimension.	
(Q16) Examination results are not recognized by any certification or selection system.	(Q16) Examination results are recognized by the certification or selection system in the country.	(Q16) Examination results are recognized by one certification or selection system in another country.	(Q16) Examination results are recognized by two or more certification or selection systems in another country.	
(Q17) The examination office does not have the required facilities to carry out the examination.	(Q17) The examination office has some of the required facilities to carry out the examination.	(Q17) The examination office has all of the required facilities to carry out the examination.	(Q17) The examination office has state-of-the-art facilities to carry out the examination.	
EC5—Having effective human resources				
(Q18) There is no staff to carry out the examination.	(Q18, Q19) The examination office is inadequately staffed to effectively carry out the examination; issues are pervasive.	(Q18, Q19) The examination office is adequately staffed to carry out the examination effectively, with minimal issues.	(Q18, Q19) The examination office is adequately staffed to carry out the assessment effectively, with no issues.	
(Q20) The country/system does not offer opportunities that prepare for work on the examination.	This option does not apply to this dimension.	(Q20) The country/system offers some opportunities that prepare for work on the examination.	(Q20) The country/system offers a wide range of opportunities that prepare for work on the examination.	

LATENT Absence of, or deviation from, the attribute	EMERGING On way to meeting minimum standard	ESTABLISHED Acceptable minimum standard	ADVANCED Best practice	Justification
System Alignment (SA) Degree to which the assessment is coherent with other components of the education system.				
SA1—Aligning examinations with learning goals and opportunities to learn				
(Q21) It is not clear what the examination measures.	This option does not apply to this dimension.	(Q21) There is a clear understanding of what the examination measures.	This option does not apply to this dimension.	
(Q22) What the examination measures is questioned by some stakeholder groups.	This option does not apply to this dimension.	(Q22) What is measured by the examination is largely accepted by stakeholder groups.	This option does not apply to this dimension.	
(Q23, Q24) Material to prepare for the examination is minimal and it is accessible to very few students.	(Q23, Q24) There is some material to prepare for the examination that is accessible to some students.	(Q23, Q24) There is comprehensive material to prepare for the examinations that is accessible to most students.	(Q23, Q24) There is comprehensive material to prepare for the examination that is accessible to all students.	
SA2—Providing teachers with opportunities to learn about the examination				
(Q25) There are no courses or workshops on examinations available to teachers.	(Q25) There are no up-to-date courses or workshops on examinations available to teachers.	(Q25) There are up-to-date voluntary courses or workshops on examinations available to teachers.	(Q25) There are up-to-date compulsory courses or workshops on examinations for teachers.	
(Q26) Teachers are excluded from all examination-related tasks.	(Q26) Teachers are involved in very few examination-related tasks.	(Q26) Teachers are involved in some examination-related tasks.	(Q26) Teachers are involved in most examination-related tasks.	
Assessment Quality (AQ) Degree to which the assessment meets quality standards, is fair, and is used in an effective way.				
AQ1—Ensuring quality				
(Q27) There is no technical report or other documentation.	(Q27) There is some documentation on the examination, but it is not in a formal report format.	(Q27) There is a comprehensive technical report but with restricted circulation.	(Q27) There is a comprehensive, high-quality technical report available to the general public.	
(Q28) There are no mechanisms in place to ensure the quality of the examination.	This option does not apply to this dimension.	(Q28) There are limited systematic mechanisms in place to ensure the quality of the examination.	(Q28) There are varied and systematic mechanisms in place to ensure the quality of the examination.	

LATENT Absence of, or deviation from, the attribute	EMERGING On way to meeting minimum standard	ESTABLISHED Acceptable minimum standard	ADVANCED Best practice	Justification
AQ2—Ensuring fairness				
(Q29) Inappropiate behavior surrounding the examination process is high.	(Q29) Inappropiate behavior surrounding the examination process is moderate.	(Q29) Inappropiate behavior surrounding the examination process is low.	(Q29) Inappropiate behavior surrounding the examination process is marginal.	
(Q30) The examination results lack credibility for all stakeholder groups.	(Q30) The examination results are credible for some stakeholder groups.	(Q30) The examination results are credible for all stakeholder groups.	This option does not apply to this dimension.	
(Q31, Q32) The majority of students (over 50%) may not take the examination because of language, gender or other equivalent barriers.	(Q31, Q32) A significant proportion of students (10%-50%) may not take the examination because of language, gender or other equivalent barriers.	(Q31, Q32) A small proportion of students (less than 10%) may not take the examination because of language, gender or other equivalent barriers.	(Q31) All students can take the examination; there are no language, gender, or other equivalent barriers.	
AQ3—Using examination information in a fair way				
(Q33) Examination results are not used in an appropriate way by all stakeholder groups.	(Q33) Examination results are used by some stakeholder groups in an appropriate way.	(Q33) Examination results are used by most stakeholder groups in an appropriate way.	(Q33) Examination results are used by all stakeholder groups in an appropriate way.	
(Q34) Student names and results are made public.	This option does not apply to this dimension.	(Q34) Student results are confidential.	This option does not apply to this dimension.	
AQ4—Ensuring positive consequences of the examination				
(Q35) There are no options for students who do not perform well on the examination, or students must leave the education system.	(Q35) There are very limited options for students who do not perform well on the examination.	(Q35) There are some options for students who do not perform well on the examination.	(Q35) There are a variety of options for students who do not perform well on the examination.	
(Q36) There are no mechanisms in place to monitor the consequences of the examination.	This option does not apply to this dimension.	(Q36) There are some mechanisms in place to monitor the consequences of the examination.	(Q36) There are a variety of mechanisms in place to monitor the consequences of the examination.	

Source: World Bank.

National Large-Scale Assessment (NLSA)

LATENT Absence of, or deviation from, the attribute	EMERGING On way to meeting minimum standard	ESTABLISHED Acceptable minimum standard	ADVANCED Best practice	Justification
Enabling Context (EC) Overall framework of policies, leadership, organizational structures, fiscal, and human resources in which NLSA activity takes place in an education system and the extent to which that framework is conducive to, or supportive of, the NLSA activity.				
EC1—Setting clear policies for NLSA				
(Q3_III) No NLSA exercise has taken place.	(Q3_III) The NLSA has been operating on an irregular basis.	(Q3_III) The NLSA is a stable program that has been operating regularly.	This option does not apply to this dimension.	
(Q5) There is no policy document pertaining to NLSA.	(Q5) There is an informal or draft policy document that authorizes the NLSA.	(Q5) There is a formal policy document that authorizes the NLSA.	This option does not apply to this dimension.	
Does not apply.	(Q7) The policy document is not available to the public.	(Q7) The policy document is available to the public.	This option does not apply to this dimension.	
(Q8) There is no plan for NLSA activity.	This option does not apply to this dimension.	(Q8, Q9) There is a general understanding that the NLSA will take place.	(Q8, Q9) There is a written NLSA plan for the coming years.	
EC2—Having strong public engagement for NLSA				
(Q11, Q12) All stakeholder groups strongly oppose the NLSA.	(Q11, Q12) Some stakeholder groups oppose the NLSA.	(Q11, Q12) Most stakeholders groups support the NLSA.	(Q11, Q12) All stakeholder groups support the NLSA.	
EC3—Having regular funding for NLSA				
(Q13) There is no funding allocated to the NLSA.	(Q13) There is irregular funding allocated to the NLSA.	(Q13) There is regular funding allocated to the NLSA.	This option does not apply to this dimension.	
Does not apply.	(Q14) Funding covers some core NLSA activities: design, administration, analysis or reporting.	(Q14) Funding covers all core NLSA activities: design, administration, analysis and reporting.	This option does not apply to this dimension.	
Does not apply.	(Q14) Funding does not cover research and development activities.	This option does not apply to this dimension.	(Q14) Funding covers research and development activities.	

LATENT Absence of, or deviation from, the attribute	EMERGING On way to meeting minimum standard	ESTABLISHED Acceptable minimum standard	ADVANCED Best practice	Justification
EC4—Having strong organizational structures for NLSA				
(Q15) There is no NLSA office, ad hoc unit or team.	(Q15) The NLSA office is a temporary agency or group of people.	(Q15) The NLSA office is a permanent agency, institution, or unit.	This option does not apply to this dimension.	
This option does not apply to this dimension.	(Q16, Q17) Political considerations regularly hamper technical considerations.	(Q16, Q17) Political considerations sometimes hamper technical considerations.	(Q16, Q17) Political considerations never hamper technical considerations.	
This option does not apply to this dimension.	(Q18, Q19) The NLSA office is not accountable to a clearly recognized body.	(Q18, Q19) The NLSA office is accountable to a clearly recognized body.	This option does not apply to this dimension.	
EC5—Having effective human resources for NLSA				
(Q20)) There is no staff allocated for running a NLSA.	(Q20, Q21) The NLSA office is inadequately staffed to effectively carry out the assessment.	(Q20, Q21) The NLSA office is adequately staffed to carry out the NLSA effectively, with minimal issues.	(Q20, Q21) The NLSA office is adequately staffed to carry out the NLSA effectively, with no issues.	
(Q22) The country/system does not offer opportunities that prepare individuals for work on NLSA.	This option does not apply to this dimension.	(Q22) The country/system offers some opportunities to prepare individuals for work on the NLSA.	(Q22) The country/system offers a wide range of opportunities to prepare individuals for work on the NLSA.	
System Alignment (SA) Degree to which the NLSA is coherent with other components of the education system.				
SA1—Aligning the NLSA with learning goals				
(Q23) It is not clear if the NLSA is based on curriculum or learning standards.	This option does not apply to this dimension.	(Q23) The NLSA measures performance against curriculum or learning standards.	This option does not apply to this dimension.	
(Q24) What the NLSA measures is generally questioned by stakeholder groups.	This option does not apply to this dimension.	(Q24) What the NLSA measures is questioned by some stakeholder groups.	(Q24) What the NLSA measures is largely accepted by stakeholder groups.	
(Q25) There are no mechanisms in place to ensure that the NLSA accurately measures what it is supposed to measure.	(Q25, Q26) There are ad hoc reviews of the NLSA to ensure that it measures what it is intended to measure.	(Q25, Q26) There are regular internal reviews of the NLSA to ensure that it measures what it is intended to measure.	This option does not apply to this dimension.	

LATENT Absence of, or deviation from, the attribute	EMERGING On way to meeting minimum standard	ESTABLISHED Acceptable minimum standard	ADVANCED Best practice	Justification
SA2—Providing teachers with opportunities to learn about the NLSA				
(Q27) There are no courses or workshops on the NLSA.	(Q27, Q28) There are occassional courses or workshops on the NLSA.	(Q27, Q28) There are some courses or workshops on the NLSA offered on a regular basis.	(Q27, Q28) There are widely available high-quality courses or workshops on the NLSA offered on a regular basis.	
Assessment Quality (AQ) Degree to which the NLSA meets technical standards, is fair, and is used in an effective way.				
AQ1—Ensuring the quality of the NLSA				
(Q29) No options are offered to include all groups of students in the NLSA.	This option does not apply to this dimension.	(Q29) At least one option is offered to include all groups of students in the NLSA.	(Q29) Different options are offered to include all groups of students in the NLSA.	
(Q30) There are no mechanisms in place to ensure the quality of the NLSA.	This option does not apply to this dimension.	(Q30) There are some mechanisms in place to ensure the quality of the NLSA.	(Q30) There are a variety of mechanisms in place to ensure the quality of the NLSA.	
(Q31) There is no technical report or other documentation about the NLSA.	(Q31) There is some documentation about the technical aspects of the NLSA, but it is not in a formal report format.	(Q31) There is a comprehensive technical report, but with restricted circulation.	(Q31) There is a comprehensive, high-quality technical report available to the general public.	
AQ2—Ensuring effective uses of the NLSA				
(Q32) NLSA results are not disseminated.	(Q32) NLSA results are poorly disseminated.	(Q32) NLSA results are disseminated in an effective way.	This option does not apply to this dimension.	
(Q33) NLSA information is not used or is used in ways inconsistent with the purposes or the technical characteristics of the assessment.	This option does not apply to this dimension.	(Q33) NLSA results are used by some stakeholder groups in a way that is consistent with the purposes and technical characteristics of the assessment.	(Q33) NLSA information is used by all stakeholder groups in a way that is consistent with the purposes and technical characteristics of the assessment.	
(Q34) There are no mechanisms in place to monitor the consequences of the NLSA.	This option does not apply to this dimension.	(Q34) There are some mechanisms in place to monitor the consequences of the NLSA.	(Q34) There are a variety of mechanisms in place to monitor the consequences of the NLSA.	

Source: World Bank.

International Large-Scale Assessment (ILSA)

LATENT Absence of, or deviation from, the attribute	EMERGING On way to meeting minimum standard	ESTABLISHED Acceptable minimum standard	ADVANCED Best practice	Justification
Enabling Context (EC) Overall framework of policies, leadership, organizational structures, fiscal and human resources in which ILSA takes place in an education system and the extent to which that framework is conducive to, or supportive of, the ILSA activity.				
EC1—Setting clear policies for ILSA				
(Q1, Q2) The country/system has not participated in an ILSA in the last 10 years.	This option does not apply to this dimension.	(Q1, Q2) The country/system has participated in at least one ILSA in the last 10 years.	(Q1, Q2) The country/system has participated in two or more ILSA in the last 10 years.	
(Q3) The country/system has not taken concrete steps to participate in an ILSA in the next 5 years.	This option does not apply to this dimension.	(Q3) The country/system has taken concrete steps to participate in at least one ILSA in the next 5 years.	This option does not apply to this dimension.	
(Q5) There is no policy document that addresses participation in ILSA.	(Q5) There is an informal or draft policy document that addresses participation in ILSA.	(Q5) There is a formal policy document that addresses participation in ILSA.	This option does not apply to this dimension.	
Does not apply.	(Q7) The policy document is not available to the public.	(Q7) The policy document is available to the public.	This option does not apply to this dimension.	
EC2—Having regular funding for ILSA				
(Q8) There is no funding for participation in ILSA.	(Q9) There is funding from loans or external donors.	(Q9) There is regular funding allocated at discretion.	(Q9) There is regular funding approved by law, decree or norm.	
This option does not apply to this dimension.	(Q10) Funding covers some core activities of the ILSA.	(Q10) Funding covers all core activities of the ILSA.	This option does not apply to this dimension.	
(Q10) Funding does not cover research and development activities.	This option does not apply to this dimension.	This option does not apply to this dimension.	(Q10) Funding covers research and development activities.	

LATENT Absence of, or deviation from, the attribute	EMERGING On way to meeting minimum standard	ESTABLISHED Acceptable minimum standard	ADVANCED Best practice	Justification
EC3—Having effective human resources for ILSA				
(Q11, Q12) There is no team or national/system coordinator to carry out the ILSA activities.	(Q11, Q12) There is a team or national/system coordinator to carry out the ILSA activities.	(Q11, Q12) There is a team and national/system coordinator to carry out the ILSA activities.	This option does not apply to this dimension.	
This option does not apply to this dimension.	(Q13) The national/system coordinator or other designated team member is not fluent in the official language of the ILSA exercise.	(Q13) The national/system coordinator is fluent in the official language of the ILSA exercise.	This option does not apply to this dimension.	
This option does not apply to this dimension.	(Q13, Q14, Q15) The ILSA office is inadequately staffed or trained to carry out the assessment effectively.	(Q13, Q14, Q15) The ILSA office is adequately staffed or trained to carry out the ILSA effectively, with minimal issues.	(Q13, Q14, Q15) The ILSA office is adequately staffed and trained to carry out the ILSA effectively, with no issues.	
System Alignment (SA) Degree to which the ILSA is coherent with other components of the education system.				
SA1—Providing opportunities to learn about ILSA				
(Q14) The ILSA team has not attended international workshops or meetings.	(Q14) The ILSA team attended some international workshops or meetings.	(Q14) The ILSA team attended all international workshops or meetings.	This option does not apply to this dimension.	
(Q16) The country/system offers no opportunities to learn about ILSA.	This option does not apply to this dimension.	(Q16, Q17) The country/system offers some opportunities to learn about ILSA.	(Q16,Q17) The country/system offers a wide range of opportunities to learn about ILSA.	
This option does not apply to this dimension.	This option does not apply to this dimension.	(Q18) Opportunities to learn about ILSA are available to the country's/system's ILSA team members only.	(Q18) Opportunities to learn about ILSA are available to a wide audience, in addition to the country's/system's ILSA team members.	

LATENT Absence of, or deviation from, the attribute	EMERGING On way to meeting minimum standard	ESTABLISHED Acceptable minimum standard	ADVANCED Best practice	Justification
Assessment Quality (AQ) Degree to which the ILSA meets technical quality standards, is fair, and is used in an effective way.				
AQ1—Ensuring the quality of ILSA				
(Q19) Data from the ILSA has not been published.	(Q19) The country/system met sufficient standards to have its data presented beneath the main display of the international report or in an annex.	(Q19) The country/system met all technical standards required to have its data presented in the main displays of the international report.	This option does not apply to this dimension.	
(Q20) The country/system has not contributed new knowledge on ILSA.	This option does not apply to this dimension.	This option does not apply to this dimension.	(Q20) The country/system has contributed new knowledge on ILSA.	
AQ2—Ensuring effective uses of ILSA				
(Q21, Q22) If any, country/system-specific results and information are not disseminated in the country/system.	(Q21, Q22) Country/system-specific results and information are disseminated irregularly in the country/system.	(Q21, Q22) Country/system-specific results and information are regularly disseminated in the country/system.	(Q21, Q22) Country/system-specific results and information are regularly and widely disseminated in the country/system.	
(Q21, Q23) Products to provide feedback to schools and educators about the ILSA results are not made available.	This option does not apply to this dimension.	(Q21, Q23) Products to provide feedback to schools and educators about the ILSA results are sometimes made available.	(Q21, Q23) Products to provide feedback to schools and educators about ILSA results are systematically made available.	
(Q24) There is no media coverage of the ILSA results.	(Q24) There is limited media coverage of the ILSA results.	(Q24) There is some media coverage of the ILSA results.	(Q24) There is wide media coverage of the ILSA results.	
(Q25, Q26) If any, country/system-specific results and information from the ILSA are not used to inform decision making in the country/system.	(Q26) Results from the ILSA are used in a limited way to inform decision making in the country/system.	(Q26) Results from the ILSA are used in some ways to inform decision making in the country/system.	(Q26) Results from the ILSA are used in a variety of ways to inform decision making in the country/system.	
(Q27) It is not clear that decisions based on ILSA results have had a positive impact on students' achievement levels.	This option does not apply to this dimension.	This option does not apply to this dimension.	(Q27) Decisions based on the ILSA results have had a positive impact on students' achievement levels.	

Source: World Bank.

Appendix 3. Example of Using the Rubrics to Evaluate a National Large-Scale Assessment Program

COUNTRY X National Large-Scale Assessment (NLSA) Rubric					Score	Adjusted Score (with Constraint)	Default Weight	Preliminary Level of Development (based on Adjusted Score)	Notes
LATENT **Absence of, or deviation from, the attribute**	**EMERGING** **On way to meeting minimum standard**	**ESTABLISHED** **Acceptable minimum standard**	**ADVANCED** **Best practice**	**JUSTIFICATION**	2.32	2.11	1	EMERGING	
Enabling Context (EC) Overall framework of policies, leadership, organizational structures, fiscal, and human resources in which NLSA activity takes place in an education system and the extent to which that framework is conducive to, or supportive of, the NLSA activity.					2.63	2	0.33	Emerging	
EC1—Setting clear policies for NLSA					2	2	0.2		
(Q3_III) No NLSA exercise has taken place.	(Q3_III) The NLSA has been operating on an irregular basis.	(Q3_III) The NLSA is a stable program that has been operating regularly.	This option does not apply to this dimension.	In 2009, the NLSA program in Country X was operating on a regular basis. However, funding for the various NLSA exercises was being sourced from different donors, and the assessments were taking place roughly every 3 to 4 years.	3		0.25		Constraint
(Q5) There is no policy document pertaining to NLSA.	(Q5) There is an informal or draft policy document that authorizes the NLSA.	(Q5) There is a formal policy document that authorizes the NLSA.	This option does not apply to this dimension.	In 2009, Country X did not have any kind (formal, informal, draft) of policy document on NLSA activity.	1		0.25		Constraint
Does not apply.	(Q7) The policy document is not available to the public.	(Q7) The policy document is available to the public.	This option does not apply to this dimension.	There was no policy document available in 2009.	1		0.25		
(Q8) There is no plan for NLSA activity.	This option does not apply to this dimension.	(Q8, Q9) There is a general understanding that the NLSA will take place.	(Q8, Q9) There is a written NLSA plan for the coming years.	Although there was no formal policy document underpinning the NLSA in 2009, there was a general understanding that the NLSA would take place every 3 to 4 years.	3		0.25		

COUNTRY X National Large-Scale Assessment (NLSA) Rubric					Score	Adjusted Score (with Constraint)	Default Weight	Preliminary Level of Development (based on Adjusted Score)	Notes
LATENT **Absence of, or deviation from, the attribute**	**EMERGING** **On way to meeting minimum standard**	**ESTABLISHED** **Acceptable minimum standard**	**ADVANCED** **Best practice**	**JUSTIFICATION**	2.32	2.11	1	EMERGING	
EC2—Having strong public engagement for NLSA					4		0.2		
(Q11, Q12) All stakeholder groups strongly oppose the NLSA.	(Q11, Q12) Some stakeholder groups oppose the NLSA.	(Q11, Q12) Most stakeholders groups support the NLSA.	(Q11, Q12) All stakeholder groups support the NLSA.	Based on our information, there is no opposition to the NLSA.	4		1		
EC3—Having regular funding for NLSA					2	2	0.2		
(Q13) There is no funding allocated to the NLSA.	(Q13) There is irregular funding allocated to the NLSA.	(Q13) There is regular funding allocated to the NLSA.	This option does not apply to this dimension.	NLSA activity is partially funded by the MOE and partially by donors. The funding is still ad hoc.	2		0.33		Constraint
Does not apply.	(Q14) Funding covers some core NLSA activities: design, administration, analysis or reporting.	(Q14) Funding covers all core NLSA activities: design, administration, analysis and reporting.	This option does not apply to this dimension.	Funding has tended to cover only basic aspects of NLSA activities. Sometimes, there has been insufficient funding to cover all core activities.	2		0.33		
Does not apply.	(Q14) Funding does not cover research and development activities.	This option does not apply to this dimension.	(Q14) Funding covers research and development activities.	Funding has primarily focused on supporting the actual carrying out of NLSA activities and not on R&D or secondary analysis.	2		0.33		
EC4—Having strong organizational structures for NLSA					2.67	2	0.2		
(Q15) There is no NLSA office, ad hoc unit or team.	(Q15) The NLSA office is a temporary agency or group of people.	(Q15) The NLSA office is a permanent agency, institution or unit.	This option does not apply to this dimension.	In 2009, the NLSA team was comprised of a small number of staff (4), some of whom had no background or training in NLSA. There was no permanent unit, and an institutional home was still being worked out.	2		0.33		Constraint
This option does not apply to this dimension.	(Q16, Q17) Political considerations regularly hamper technical considerations.	(Q16, Q17) Political considerations sometimes hamper technical considerations.	(Q16, Q17) Political considerations never hamper technical considerations.	There is no precedence of political considerations hampering technical considerations.	4		0.33		

COUNTRY X National Large-Scale Assessment (NLSA) Rubric					Score	Adjusted Score (with Constraint)	Default Weight	Preliminary Level of Development (based on Adjusted Score)	Notes
LATENT **Absence of, or deviation from, the attribute**	**EMERGING** **On way to meeting minimum standard**	**ESTABLISHED** **Acceptable minimum standard**	**ADVANCED** **Best practice**	**JUSTIFICATION**	2.32	2.11	1	EMERGING	
This option does not apply to this dimension.	(Q18, Q19) The NLSA office is not accountable to a clearly recognized body.	(Q18, Q19) The NLSA office is accountable to a clearly recognized body.	This option does not apply to this dimension.	In 2009, the NLSA office was not accountable to a clearly recognized body. This is because it was in transition from one institutional home to another.	2		0.33		
EC5—Having effective human resources for NLSA					2.5		0.2		
(Q20)) There is no staff allocated for running a NLSA.	(Q20, Q21) The NLSA office is inadequately staffed to effectively carry out the assessment.	(Q20, Q21) The NLSA office is adequately staffed to carry out the NLSA effectively, with minimal issues.	(Q20, Q21) The NLSA office is adequately staffed to carry out the NLSA effectively, with no issues.	In 2009, the NLSA office did not have sufficient staff to effectively carry out NLSA activities.	2		0.5		
(Q22) The country/system does not offer opportunities that prepare individuals for work on NLSA.	This option does not apply to this dimension.	(Q22) The country/system offers some opportunities to prepare individuals for work on the NLSA.	(Q22) The country/system offers a wide range of opportunities to prepare individuals for work on the NLSA.	There were some large-scale assessment- and measurement-related courses offered by the main university in Country X.	3		0.5		
System Alignment (SA) Degree to which the NLSA is coherent with other components of the education system.					2		0.33	Emerging	
SA1—Aligning the NLSA with learning goals					3		0.5		
(Q23) It is not clear if the NLSA is based on curriculum or learning standards.	This option does not apply to this dimension.	(Q23) The NLSA measures performance against curriculum or learning standards.	This option does not apply to this dimension.	The NLSA was aligned with existing curriculum and standards.	3		0.33		
(Q24) What the NLSA measures is generally questioned by stakeholder groups.	This option does not apply to this dimension.	(Q24) What the NLSA measures is questioned by some stakeholder groups.	(Q24) What the NLSA measures is largely accepted by stakeholder groups.	The MOE and other stakeholders have accepted the NLSA.	4		0.33		
(Q25) There are no mechanisms in place to ensure that the NLSA accurately measures what it is supposed to measure.	(Q25, Q26) There are ad hoc reviews of the NLSA to ensure that it measures what it is intended to measure.	(Q25, Q26) There are regular internal reviews of the NLSA to ensure that it measures what it is intended to measure.	This option does not apply to this dimension.	In 2009, there were some procedures in place for reviewing the alignment of the NLSA test with the constructs/content it was intended to measure, but these procedures were not formalized or standardized.	2		0.33		

COUNTRY X National Large-Scale Assessment (NLSA) Rubric					Score	Adjusted Score (with Constraint)	Default Weight	Preliminary Level of Development (based on Adjusted Score)	Notes
LATENT **Absence of, or deviation from, the attribute**	**EMERGING** **On way to meeting minimum standard**	**ESTABLISHED** **Acceptable minimum standard**	**ADVANCED** **Best practice**	**JUSTIFICATION**	2.32	2.11	1	EMERGING	
SA2—Providing teachers with opportunities to learn about the NLSA					1		0.5		
(Q27) There are no courses or workshops on the NLSA.	(Q27, Q28) There are occassional courses or workshops on the NLSA.	(Q27, Q28) There are some courses or workshops on the NLSA offered on a regular basis.	(Q27, Q28) There are widely available high-quality courses or workshops on the NLSA offered on a regular basis.	The only courses or workshops associated with previous NLSA exercises have been for policymakers and high-level educators, and not for classroom teachers.	1		1		
Assessment Quality (AQ) Degree to which the NLSA meets technical standards, is fair and is used in an effective way.					2.33		0.33	Emerging or Established	
AQ1—Ensuring the quality of the NLSA					2.67		0.5		
(Q29) No options are offered to include all groups of students in the NLSA.	This option does not apply to this dimension.	(Q29) At least one option is offered to include all groups of students in the NLSA.	(Q29) Different options are offered to include all groups of students in the NLSA.	The NLSA is translated into the language of instruction for each region.	3		0.33		
(Q30) There are no mechanisms in place to ensure the quality of the NLSA.	This option does not apply to this dimension.	(Q30) There are some mechanisms in place to ensure the quality of the NLSA.	(Q30) There are a variety of mechanisms in place to ensure the quality of the NLSA.	In 2009, there were some procedures in place for reviewing the alignment of the NLSA test with the constructs/content it was intended to measure. This would allow us to say that there were 'some mechanisms in place to ensure the quality of the NLSA.'	3		0.33		
(Q31) There is no technical report or other documentation about the NLSA.	(Q31) There is some documentation about the technical aspects of the NLSA, but it is not in a formal report format.	(Q31) There is a comprehensive technical report, but with restricted circulation.	(Q31) There is a comprehensive, high-quality technical report available to the general public.	In 2009, no formal technical reports were available for the NLSA.	2		0.33		
AQ2—Ensuring effective uses of the NLSA					2		0.5		
(Q32) NLSA results are not disseminated.	(Q32) NLSA results are poorly disseminated.	(Q32) NLSA results are disseminated in an effective way.	This option does not apply to this dimension.	In 2009, NLSA results were not being widely disseminated to key stakeholders. Few copies of the report were available.	2		0.33		

COUNTRY X National Large-Scale Assessment (NLSA) Rubric					Score	Adjusted Score (with Constraint)	Default Weight	Preliminary Level of Development (based on Adjusted Score)	Notes
LATENT **Absence of, or deviation from, the attribute**	**EMERGING** **On way to meeting minimum standard**	**ESTABLISHED** **Acceptable minimum standard**	**ADVANCED** **Best practice**	**JUSTIFICATION**	2.32	2.11	1	EMERGING	
(Q33) NLSA information is not used or is used in ways inconsistent with the purposes or the technical characteristics of the assessment.	This option does not apply to this dimension.	(Q33) NLSA results are used by some stakeholder groups in a way that is consistent with the purposes and technical characteristics of the assessment.	(Q33) NLSA information is used by all stakeholder groups in a way that is consistent with the purposes and technical characteristics of the assessment.	In 2009, the NLSA results were, to a certain extent, used for curriculum development and teacher training.	3		0.33		
(Q34) There are no mechanisms in place to monitor the consequences of the NLSA.	This option does not apply to this dimension.	(Q34) There are some mechanisms in place to monitor the consequences of the NLSA.	(Q34) There are a variety of mechanisms in place to monitor the consequences of the NLSA.	In 2009, there were no mechanisms in place to monitor the consequences of the NLSA.	1		0.33		

Source: World Bank.