Module 3: Developing Monitoring and Evaluation Framework

3.2 Monitoring the quality of learning and learning outcomes

This reading material is reproduced for training purpose and extracted from:


The full publications can be accessed at:
What Matters Most for Student Assessment Systems: A Framework Paper

Marguerite Clarke

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

Introduction

Assessment is the process² 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

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.
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 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). 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.

---

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.
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. 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). 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).

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. Black and Wiliam (1998) conclude:

---

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.

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
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. 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 suggest that simply reporting information about average school scores on these assessments also can

---

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).
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 testing11

---

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.
• 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

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

---

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 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.
large-scale, system-level assessments for monitoring and providing policymaker- 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 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 and

---

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.
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. 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 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).

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,

---

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.

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.
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). 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
- 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

---

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).
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; 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

---

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.
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 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.
Appendix 1. Assessment Types and Their Key Differences

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

Toward Universal Learning
What Every Child Should Learn

Executive Summary
Introduction

The benefits of education—for national development, individual prosperity, health and social stability—are well known, but for these benefits to accrue children in school have to be learning. Despite commitments and progress in improving access to education at the global level, including Millennium Development Goal (MDG) 2 on universal primary education and the Education for All (EFA) Goals, levels of learning are still too low. According to estimates in the 2012 EFA Global Monitoring Report, at least 250 million primary-school-age children around the world are not able to read, write or count well enough to meet minimum learning standards, including those children who have spent at least four years in school (UNESCO 2012). Worse still, we may not know the full scale of the crisis and this figure is likely to be an underestimate because measurement of learning outcomes among children and youth is limited and, relative to the measurement of access, more difficult to assess at the global level.

To advance progress for children and youth around the world, it is critical that learning is recognized as essential for human development. As EFA and the MDGs sunset in 2015, and the UN Secretary-General promotes the Global Education First initiative, the education sector has a unique window of opportunity to raise awareness of international education goals and ensure that learning becomes a central component of the global development agenda. To do this, the global education community must work together to define global ambition on improving learning and propose practical actions to deliver and measure progress.

In response to this need, UNESCO through its Institute for Statistics (UIS) and the Center for Universal Education (CUE) at the Brookings Institution have co-convened the Learning Metrics Task Force (LMTF) project. The overarching objective of the project is to catalyze a shift in the global conversation on education from a focus on access to access plus learning. Based on recommendations of technical working groups and input from broad global consultations, the task force aims to make recommendations to help countries and international organizations measure and improve learning outcomes for children and youth worldwide.

The Process

With members representing national and regional governments, EFA-convening agencies, regional political bodies, civil society, and donor agencies¹, the task force is engaged in an 18-month-long process to build consensus around three essential questions addressed in the following order:

- Phase I: What learning is important for all children and youth?
- Phase II: How should learning outcomes be measured?
- Phase III: How can measurement of learning improve education quality?

From the outset, the task force agreed on a set of basic principles to guide its proceedings:

- The work of the task force should be open, transparent and inclusive, with balanced representation from the global north and south.
- Rather than focusing solely on the developing world, task force recommendations should be truly global, addressing learning in all countries.
- Equity within countries should be emphasized in ad-

¹For a list of task force members, working group members, and consultation participants, see the full report.
dition to overall national learning levels, with a particular attention to marginalized groups.

• The recommendations of the task force should not be limited to current capacity for measurement, but should look ahead to the next 15 years, allowing for changing needs and future innovations in technology and assessment.

During each phase of the project, a working group of technical experts, academics and practitioners around the world works collaboratively to investigate existing policies related to measuring learning, review the research, and analyze feedback from global consultations. The task force then makes decisions based on recommendations from the working group, and releases a report with its findings. This report presents the results of Phase I.

The Purpose of the Report

For Phase I, the Standards Working Group was charged with investigating whether certain competencies, knowledge or areas of learning are important for all children and youth to master in order to succeed in school and life. The primary purpose of this report is to document the Phase I process and present the rationale for the learning domains framework proposed by the task force. Subsequent reports, to be released later in 2013, will build on this foundation by providing actionable recommendations for stakeholders in the global education community.

Before identifying what learning is important, the working group first needed to examine the various contexts in which children are learning around the world, from early childhood (birth through primary school entry) through primary school and postprimary (end of primary through end of lower secondary) levels.

When and Where Children Learn

Early Childhood

Globally, 164 million children are enrolled in preschool programs, and the preprimary gross enrollment ratio (GER) is 48 percent (UNESCO 2012). However, access to preprimary programs is unevenly distributed — in low-income countries the GER for preprimary is only 15 percent. The children least likely to be enrolled in preschool are those belonging to minority ethnic groups, those with less educated mothers, and those who speak a home language different from the language used in school (UNESCO 2012). These are also the children who are most likely to benefit from high-quality preprimary programs.

While many children, especially in high-income countries, attend formal, regulated preprimary programs, the majority of the world’s young children learn in nonformal contexts through unstructured or informal processes. For these children, learning typically occurs in the home and community through interactions with parents, siblings and other family members. Even when children are enrolled in preprimary programs, they may not be exposed to high-quality formal early learning opportunities.

Primary

Partly as a result of the push for universal primary education, the majority (89 percent) of primary age children are now enrolled in school (UNESCO 2012). Free, compulsory primary education is recognized as a fundamental human right (United Nations 1948), and primary education is compulsory in almost every
country (UNESCO Institute for Statistics [UIS] 2012). Still, there are nearly 61 million out-of-school children of primary-school age, a number that has stagnated since 2008 (UNESCO 2012).

While some children are either not enrolled in school or are enrolled in nonformal programs, the majority of children globally are learning in formal contexts. However, the degree to which formal processes are good enough to ensure children’s right to a decent education depends in large part on the quality of the teachers, curriculum and materials found in the school. In schools where there are enough qualified teachers and materials to respond to each individual child’s learning needs, academic learning occurs through formal processes. In schools where teachers are not properly qualified, are overextended or do not come to work regularly, learning still occurs through peer-to-peer interactions—but not necessarily the types of learning intended by the school system (Wagner et al. 2012).

Postprimary

The category of postprimary refers to the various contexts in which children learn beyond primary schooling. For most children, “postprimary” refers to secondary education. Given the diverse areas of specialization students engage in after secondary school, the task force decided to limit its recommendations at the postprimary level to lower secondary. The UIS reports that in 2010, lower secondary education was part of compulsory education in three out of four countries reporting data, and upper secondary was included in compulsory education in approximately one out of four countries (UIS 2012). It is estimated that globally, 91 percent of children who entered school stay there until the end of primary school, and 95 percent of those students transition to secondary school. However, for children in low-income countries, only 59 percent make it to the last year of primary school, and 72 percent of those students successfully transition to secondary school (UIS 2012). For children who do not attend secondary school, learning occurs mainly through work, family and community experiences (i.e., nonformal, unstructured contexts) (Wagner et al. 2012).

Proposed Framework: Seven Domains of Learning

Given the diversity of structures, places, and times at which children and youth learn, it is a challenge to define what outcomes related to learning are important, especially at a global level. Furthermore, to develop a framework that would be relevant for the next 15 years, the task force recognized that it would have to take a step back from what is measurable today and consider first what learning is important for the 21st century. Feedback from interviews with key stakeholders and global consultations points to a growing demand globally for measuring learning in multiple areas, not just literacy and numeracy. Accordingly, the task force proposes a broad definition of learning that encompasses seven domains, with corresponding subdomains, as important for all children and youth to develop (see Table 1 and Figure 1).
<table>
<thead>
<tr>
<th>Domain</th>
<th>Description</th>
<th>Subdomain Examples*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical well-being</td>
<td>How children and youth use their bodies, develop motor control, and understand and exhibit appropriate nutrition, exercise, hygiene and safety practices.</td>
<td>• Physical health and hygiene</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Food and nutrition</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Physical activity</td>
</tr>
<tr>
<td>Social and emotional</td>
<td>How children and youth foster and maintain relationships with adults and peers. Also, how they perceive themselves in relation to others.</td>
<td>• Social and community values</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Civic values</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Mental health and well-being</td>
</tr>
<tr>
<td>Culture and the arts</td>
<td>Creative expression, including activities from the areas of music, theater, dance or creative movement, and the visual, media and literary arts. Also, cultural experiences in families, school, community and country.</td>
<td>• Creative arts</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Cultural knowledge</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Self- and community identity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Awareness of and respect for diversity</td>
</tr>
<tr>
<td>Literacy and communication</td>
<td>Communication in the primary language(s) of the society in which children and youth live, including speaking, listening, reading, writing, and understanding the spoken and written word in various media.</td>
<td>• Speaking and listening</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Vocabulary</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Writing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Reading</td>
</tr>
<tr>
<td>Learning approaches and cognition</td>
<td>Learning approaches describe a learners' engagement, motivation and participation in learning. Cognition is the mental process of acquiring learning through these various approaches.</td>
<td>• Persistence and attention</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Cooperation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Problem solving</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Self-direction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Critical thinking</td>
</tr>
<tr>
<td>Numeracy and mathematics</td>
<td>The science of numbers and quantitative language used universally to represent phenomena observed in the environment.</td>
<td>• Number concepts and operations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Geometry and patterns</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Mathematics application</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Data and statistics</td>
</tr>
<tr>
<td>Science and technology</td>
<td>Science is specific knowledge or a body or system of knowledge covering physical laws and general truths. Technology refers to the creation and usage of tools to solve problems.</td>
<td>• Scientific inquiry</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Life science</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Physical science</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Earth science</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Awareness and use of digital technology</td>
</tr>
</tbody>
</table>

*Subdomains listed here are by way of example only. See the main report for the full list of subdomains across each level (early childhood, primary and postprimary).
This holistic framework of learning domains was developed by drawing on:

• Existing global policies and dialogues, such as EFA and the UN Convention on the Rights of the Child, which mandate a broad definition of education and learning.

• Research supporting the importance of learning in these domains for human development, economic growth and prosperity.

• Results from global public consultation, in which more than 500 individuals in 57 countries provided feedback. The overwhelming majority of participants in the global consultation, especially those from the Global South, argued for a broad definition of learning that goes beyond basic literacy and numeracy.

Note: This framework is intended for the purpose of the Learning Metrics Task Force to identify areas in which to measure learning outcomes. It is not intended to be used as a framework for policymaking, curriculum or instruction.

Figure 1: A Global Framework of Learning Domains

Each arrow in Figure 1 represents one domain of learning, radiating outward as a child expands his or her development or competency in a given area. The half circles represent three stages in which the task force will concentrate its recommendations: early childhood (birth through primary school entry); primary and postprimary (end of primary through end of lower secondary). The arrows extend outward beyond the diagram to indicate that an individual may continue learning more deeply in a given domain at the upper secondary, tertiary, or technical/vocational level or through nonformal learning opportunities.

---

2Each arrow in Figure 1 represents one domain of learning, radiating outward as a child expands his or her development or competency in a given area. The half circles represent three stages in which the task force will concentrate its recommendations: early childhood (birth through primary school entry); primary and postprimary (end of primary through end of lower secondary). The arrows extend outward beyond the diagram to indicate that an individual may continue learning more deeply in a given domain at the upper secondary, tertiary, or technical/vocational level or through nonformal learning opportunities.
Considerations Related to Equity

The task force noted several considerations for the following populations and contexts related to the seven learning domains.

Children with Disabilities
An estimated 15 to 20 percent of students worldwide have special learning needs, and children with disabilities are less likely to enroll in and complete school than their nondisabled peers (World Health Organization and World Bank 2011). In low-income countries, their exclusion from education can be very significant and result in lifelong discrimination.

The learning domains framework covers a broad set of learning outcomes, allowing children who struggle with traditional academic or cognitive tasks to have an opportunity to demonstrate strengths in a variety of domains. With targeted instructional support and accommodations, children with disabilities can make progress toward learning goals in all seven domains. When assessing learning for children with disabilities, as with all children, a focus on individual progress can be more relevant in measuring and improving learning outcomes than a focus on absolute learning levels. More frequent and fine-grained monitoring of progress may be necessary to capture improvements in learning for children with disabilities.

Gender
Gender may be more important in discussing the determinants of learning in the classroom than in making choices about outcome measures. Gender issues are important across all domains, but especially in the domains of physical well-being, social and emotional, and learning approaches and cognition. For example, under physical well-being, the fact that girls can get pregnant and boys cannot, compounded with a social and cultural context of male power and female subservience, make necessary learning outcomes in this area quite different for boys and girls.

There is an implicit assumption in this framework that as the arrows radiate out, from level to level, children are developing and learning at a similar and steady rate. However, in many settings this is not always the case given delayed school entry ages as well as repetition rates. Thus particularly when looking at the physical well-being domain and the social and emotional domain, one needs to recognize that physical and emotional development may also be affected by age as well as by level. This is compounded by the fact that girls tend to reach puberty about two years before boys do. While one can reasonably assume that all postprimary students are older adolescents or young adults, one cannot assume that all primary students are preadolescent.

Learning in Conflict and Emergency Contexts
War and natural disasters can significantly disrupt a child’s education and learning trajectory. When children are displaced due to these circumstances, they often are excluded from school for years, sometimes even generations. However, a high-quality education in emergency situations can provide physical, psychosocial and cognitive protection that can sustain and save lives (Inter-Agency Network for Education in Emergencies [INEE] 2010). In the physical well-being and social and emotional domains, education can provide children with critical survival skills and coping
mechanisms through learning about landmine safety, HIV/AIDS prevention and conflict resolution strategies, for example. During conflict and emergencies, learning may occur in formal schooling settings, but very often it occurs in informal ways. Therefore, efforts to assess children’s learning must take into account where school-age children are, what is being taught, mother tongue and language of instruction, and a variety of other factors (INEE 2010).

Countries Demonstrating Low Levels of Learning

Currently, international capacity for measuring learning is concentrated most strongly in the domains of literacy and communication, numeracy and mathematics, and science and technology. While these measures do not provide a complete picture of what children and youth have learned, they form the basis for analysis of learning levels globally. Beatty and Pritchett (2012) argue that any learning goals proposed as part of the post-2015 development agenda should be “based on feasibility, not wishful thinking.” Goals are only successful in accelerating progress if they are perceived as achievable. In many developing countries, learning progress in the areas of literacy, mathematics and science is stagnant or even declining based on results from national and international assessments. The authors estimate that given current trends, it would take Colombia 30 years and Turkey 194 years to reach mean Organization for Economic Cooperation and Development (OECD) levels of learning as measured by Trends in International Mathematics and Science Study (TIMSS). They also posit that countries such as Indonesia, Iran, Jordan, Malaysia, Thailand and Tunisia will never catch up given current trends, as learning levels have actually declined from one testing period to the next. Among countries participating in the SACMEQ (Anglophone countries in Southern and Eastern Africa), it could take four to five generations (150 years, on average) to catch up to mean OECD learning levels in reading, given current trends.

In another report, Pritchett and Beatty (2012) find that having an overambitious curriculum in countries where achievement levels are low can lead to a “curriculum gap,” whereby more children are excluded from learning and never catch up. These countries end up being even farther behind than ones in which the curriculum is appropriate for children’s learning levels. Given these complexities, it appears that setting one-size-fits-all standards is unlikely to be useful at a global level. The challenge for the task force is to determine whether a framework can be developed that allows countries to set achievable goals based on current learning levels, understanding that a tiered system could send a message that high standards are achievable by some children and youth but not others.

Remaining Issues

The task force identified the following issues as requiring further investigation by subsequent working groups.

Should global learning goals be measured in an internationally comparable way?

The task force felt that more analysis is needed on how internationally comparable tests can influence policy and practice. Investing time and resources in internationally comparable tests only to end up at the bottom of a league table is discouraging to education ministries and may not provide the type of information necessary to improve learning levels. However, inter-
nationally comparable assessments have successfully drawn attention to gaps in curriculum and instruction and have been used to design school reform efforts in many countries. It is clear that internationally comparable assessments are useful in some contexts and less useful in others. The LMTF is interested in a tiered model of measuring learning that takes into account internationally comparable assessments in some contexts and alternative assessments in others.

Should learning assessment focus on children and youth in schools or all children and youth, regardless of where they are learning?

Given that schools are the primary vehicles for improving learning outcomes, some argue that learning assessments should be conducted only within schools to simplify and focus on making improvements to the system. Others cited low enrollment numbers in primary programs (48 percent GER globally) and secondary school (70 percent GER globally) (UNESCO 2012), especially in low- and middle-income countries, as reasons why the recommendations must extend to children outside formal school settings. This is an issue for which the answer may vary by country context—countries with universal or near-universal enrollment may compile accurate assessments of learning through schools, while countries with lower levels of enrollment may need an alternative strategy for learning assessment, such as household surveys.

Should learning be measured by age cohort or grade level?

Some argue that an age-based model would keep governments accountable for the learning of all children, whether or not they are enrolled in school. Children would need to be enrolled in schools, progressing through the levels, and learning as they go in order to meet any national or global education goals based on age cohorts (Pritchett and Beatty 2012). Others argue that the varying ages at which children begin school globally would make grade levels a fairer way of measuring learning, especially in any internationally comparable way.

Next Steps

This report documents Phase I of the LMTF project. It describes the research and policies the Standards Working Group deemed most relevant, but it is by no means a comprehensive report of education policy and learning research.

In Phase II of the project, the Measures and Methods Working Group will investigate the feasibility of measuring learning in the seven domains, taking into account current initiatives to measure learning at the local, national, regional and international levels. The second technical working group will also make recommendations for expanding the capacity for measuring learning in domains that are not currently measured on a large scale.

During the third and final phase of the initiative, the Implementation Working Group will develop recommendations for how learning assessment can be implemented to improve policy and ultimately learning outcomes. A final report with recommendations is currently scheduled for release in September 2013. Updates will continue to be available online at www.brookings.edu/learningmetrics.
Conclusion

The human right to education cannot be achieved simply by ensuring children attend school; they must also be learning while they are there. Setting goals and measuring progress have the potential to accelerate learning at the global level and building consensus around these goals and measures for learning is a crucial step toward ensuring a worldwide focus on access plus learning. The global framework of learning domains represents the task force’s vision for what every child everywhere should learn and be able to do, whether at the classroom, system, or global level, by the time they reach postprimary age.

The Learning Metrics Task Force was convened to provide a forum for all interested stakeholders to work collaboratively and share their expertise and ideas for what learning is important and how it can be measured to improve education quality. By identifying areas of consensus and discussing areas of disagreement, the task force aims to propose a framework for measuring learning that is acceptable to all stakeholders, even if it is not “ideal” for everyone. The task force recognizes that not all of the domains are feasible for a potential global learning goal; however, all seven domains are deemed important and should remain the aspiration for every child throughout the education lifespan. The second and third technical working groups will continue to refer back to the global framework of learning domains and develop a rationale for why a particular domain of learning is or is not included in subsequent recommendations.