



United Nations  
Educational, Scientific and  
Cultural Organization



Global  
Education  
Monitoring  
Report

Background paper prepared for the 2016 Global Education Monitoring Report

*Education for people and planet: Creating sustainable futures for all*

# Measures that Matter: Learning Outcome Targets for Sustainable Development Goal 4 - An examination of national, regional and international learning assessments

*This paper was commissioned by the Global Education Monitoring Report as background information to assist in drafting the 2016 report. It has not been edited by the team. The views and opinions expressed in this paper are those of the author(s) and should not be attributed to the Global Education Monitoring Report or to UNESCO. The papers can be cited with the following reference: "Paper commissioned for the Global Education Monitoring Report 2016, Education for people and planet: Creating sustainable futures for all". For further information, please contact [gemreport@unesco.org](mailto:gemreport@unesco.org).*

## 1. Table of Contents

1. Introduction.....	4
2. Four types of assessments: who, what, why, when where, how.....	8
3. International large-scale assessments .....	13
4. Who, what, why, when, where and how of existing international large-scale assessments .....	15
5. Main advantages and disadvantages of international and regional assessments.....	17
6. What is monitored: Comparison of performance standards .....	21
7. How students are assessed.....	24
8. Why countries do not already participate in international large-scale assessments .....	27
9. Use of international large-scale assessments .....	29
10. Conclusions and lessons learned .....	31

## Acronyms for Major Large-Scale International Assessments

Assessment	Name	Sponsor
ASER	Annual Status of Education Report (India and Pakistan)	Pratham
CivEd		IEA
ECES	Early Childhood Education Study	IEA
EGMA	Early Grade Mathematics Assessment	USAID/RTI
EGRA	Early Grade Reading Assessment	USAID/RTI
ERCE	Estudio Regional Comparativo y Explicativo	LLECE
ICILS	International Computer and Information Literacy Study	IEA
LaNA	Literacy and Numeracy Assessment	IEA
LAMP	Literacy Assessment and Monitoring Program	UNESCO/UIS
PASEC	Programme d'Analyse des Systemes Educatif de CONFEMEN	CONFEMEN
PIAAC	Program for the International Assessment of Adult Competencies	OECD
PIRLS	Progress in Reading Literacy Study	IEA
PIRLS Literacy	Progress in Reading Literacy Study	IEA
PISA	Programme in Student Assessment	OECD
PISA-D	PISA for Development	OECD
SACMEQ	Southern Africa C on Measurement of Education Quality	SACMEQ
STEP	Skills Towards Employability and Productivity	World Bank
TIMSS	Trends in Mathematics and Science Study	IEA
UWEZO	Uwezo means "capability" in Kiswahili	UWEZO

# 1 Introduction

This paper assesses the merits of different learning assessments for measuring progress toward the Sustainable Development Goal (SDG) for education and its associated targets, as adopted by the international community in Incheon, Republic of Korea during the World Education Forum in May, 2015 and ratified by the UN in September, 2015. It serves as a background paper for the Global Education Monitoring Report (or GEM Report),<sup>1</sup> which is an annual report that will monitor progress towards this global education goal and targets.

Improving the provision of good quality education has been an explicit aim of global education policies since the World Conference on Education for All in Jomtien, Thailand in 1990. And yet, in practice, country reforms and international aid flows have until recently mainly focused on increasing access to primary and secondary education.

This is changing in many ways. In the past, the measurement of education quality largely centered on identifying students with the skills and motivations to continue on to successive levels of education. Examinations and academic Olympiads played major roles in such identification.<sup>2</sup> Following Jomtien, multilateral donors and agencies began to provide training to national agencies to help them build the capacity for carrying out national assessments, and the number of countries implementing national assessments has increased sharply over the ensuing decades (Murphy et al, 1994; Ross & Mahlck 1990; Benevot & Koseleki 2015). Measurement of adult literacy has moved from reliance on self reported literacy to the measurement of literacy (and in some cases numeracy) via one-on-one testing of basic skills within, through citizen-led initiatives and household surveys. Basic skills among children and youth have been directly measured in early primary grades and at home, regardless of their school enrolment or attendance status. And over the past decade, large-scale regional and international assessments of reading, mathematics and science have incorporated increasing numbers of participating countries (Lockheed 2015).

These international, regional and national learning assessments strategies vary in the grades or age levels tested, target populations, coverage of the target population (sample or census), content and cognitive domains covered, types of background data gathered, and the frequency with which they are administered. They also vary in how the assessments are scored, reported and used.

---

<sup>1</sup> Formerly known as the Education for All Global Monitoring Report (GMR)

<sup>2</sup> As coordinated through such cross-national organizations as the Caribbean Examinations Council, West African Examinations Council, Cambridge Examination Board, International Baccalaureat.

Goal 4 of the SDGs pertains to education: *Ensure inclusive and quality education for all and promote lifelong learning*; it has ten associated targets related to various populations and educational levels (Box 1). At its meeting on November 2, 2015, the Interagency and Expert Group on Sustainable Development Goal indicators (IAEG-SDG) agreed on broad indicators for four of the Goal 4 targets related to education quality and learning outcomes.<sup>3</sup> It deferred discussion of indicators for one other target related to learning outcomes<sup>4</sup>. Indicators for the remaining Goal 4 targets<sup>5</sup> address participation and /or inputs rather than learning outcomes.

### **Box 1: Sustainable Development Goal 4 targets**

- 4.1. By 2030, ensure that all girls and boys complete free, equitable and quality primary and secondary education leading to relevant and effective learning outcomes
- 4.2. By 2030, ensure that all girls and boys have access to quality early childhood development, care and preprimary education so that they are ready for primary education
- 4.4.3. By 2030, ensure equal access for all women and men to affordable and quality technical, vocational and tertiary education, including university
- 4.4. By 2030, substantially increase the number of youth and adults who have relevant skills, including technical and vocational skills, for employment, decent jobs and entrepreneurship
- 4.5. By 2030, eliminate gender disparities in education and ensure equal access to all levels of education and vocational training for the vulnerable, including persons with disabilities, indigenous peoples and children in vulnerable situations
- 4.6. By 2030, ensure that all youth and a substantial proportion of adults, both men and women, achieve literacy and numeracy
- 4.7. By 2030, ensure that all learners acquire the knowledge and skills needed to promote sustainable development, including, among others, through education for sustainable development and sustainable lifestyles, human rights, gender equality, promotion of a culture of peace and non-violence, global citizenship and appreciation of cultural diversity and of culture's contribution to sustainable development
- 4.a. Build and upgrade education facilities that are child, disability and gender sensitive and provide safe, nonviolent, inclusive and effective learning environments for all
- 4.b. By 2020, substantially expand globally the number of scholarships available to developing countries, in particular least developed countries, small island developing States and African countries, for enrolment in higher education, including vocational training and information and communications technology, technical, engineering and scientific programmes, in developed countries and other developing countries
- 4.c. By 2030, substantially increase the supply of qualified teachers, including through international cooperation for teacher training in developing countries, especially least developed countries and small island developing states

<sup>3</sup> Targets 4.1, 4.2, 4.4, and 4.6; learning indicators for all 4.2 targets were not fully agreed at that meeting.

<sup>4</sup> Targets 4.7; the Technical Advisory Group proposed indicators for this target, but full discussion was deferred.

<sup>5</sup> Targets 4.3, 4.5, 4.a, 4.b, 4.c


Table 1 summarizes the learning outcome targets for different levels of education. The IAEG did not recommend any specific measures for any of the indicators. This paper reviews various assessment strategies for measuring learning outcomes relevant to these education levels and targets.

**Table 1. SDG Targets and Indicators for Learning Outcomes, by Education Level**

Education Level	Target	Indicator
Preschool	4.2 Readiness for primary education	% of children under 5 years of age who are developmentally on track in ... learning
End of Primary	4.1 Relevant and effective learning outcomes	% of children/young people at the end of [primary education] achieving at least a minimum proficiency level in (a) reading and (b) mathematics
Lower Secondary	4.1 Relevant and effective learning outcomes	% of children/young people at the end of [lower secondary education] achieving at least a minimum proficiency level in (a) reading and (b) mathematics*
Upper Secondary	4.1 No learning target	
TVET	4.4 Relevant skills, including technical and vocational skills, for employment	% of youth/adults with a minimum level of proficiency in digital literacy skills** % of youth/adults with ICT skills
Higher education	4.3 No learning target	
Youth and adults	4.6 Literacy and numeracy, skills for employment	% of the population in a given age group achieving at least a fixed level of proficiency in functional (a) literacy and (b) numeracy skills
All learners	4.7 Knowledge and skills for sustainable development	% of 15-year-old students demonstrating a fixed level of knowledge ...in environmental science and geoscience

Note: \*The Technical Advisory Group Proposal specifies “lower secondary” level, whereas the IAEG report does not. \*\*the TAG proposal includes both indicators, whereas the IAEG-SDG report includes only “ITC skills”

The agreed indicators for the various targets are somewhat narrower than the targets themselves, in two ways. First, targets 4.1, 4.2, and 4.4 call for “relevant and effective” learning outcomes, while the indicators focus on reading, mathematics, and ICT. Thus, other “relevant” learning outcomes are not included in the proposed list of indicators. Consensus about the “relevance” of various learning outcomes may be difficult to achieve,



although it appears that skills in reading and mathematics are widely accepted as universally relevant.

Second, the broad indicator for target 4.1 as specified in the IAEG-SDG report refers to the “percentage of children/young people at the end of each level of education” whereas the TAG report refers to the “percentage of children/young people (i) at the end of primary and (ii) at end of lower secondary”. Thus, the end of secondary level is not included in the education levels. For which indicators have been proposed.

## 2 Four types of assessments: who, what, why, when where, how

The literature on education assessments identifies four broad types of assessments (classroom assessments, examinations, national assessments, and large-scale international/regional assessments) and four major purposes of assessment (improving teaching, system monitoring and evaluation, student selection and student certification) for which the assessments are designed. The types and purposes are generally aligned as follows:

- classroom assessments for improving teaching<sup>6</sup>
- examinations for selection and certification of individual students
- national assessments for within-country monitoring and evaluation
- international/regional assessments for cross-national monitoring and evaluation<sup>7</sup>

For a country's assessment system to be fully effective, these various types of assessment need to be aligned with each other and with the country's education standards and policies (Clarke 2012). However, for the purpose of monitoring learning across time and across countries – that is, for the global monitoring of learning achievement – assessments must be aligned across countries. Alignment across countries requires a high degree of cross-national standardization, which is currently available only in international large-scale assessments (box 2.)

Standardization is needed both for cross-national comparability and – most importantly – for measuring change. As various psychometricians have noted: “If you want to measure change, don't change the measure.” This observation, attributed to Otis Dudley Duncan in 1969 and applied to national assessments of learning by Albert Beaton in 1988, is central to any discussion of monitoring learning achievement.

---

<sup>6</sup> Classroom assessments (including both “formative” and “summative” assessments) provide teachers information about their own students' specific learning needs and are, of course, only one way to improve teaching.

<sup>7</sup> Some countries refer to international and regional assessments for monitoring their national progress, and a few countries (e.g. Brazil, Canada) draw sufficiently large samples for international assessments to be used in monitoring learning across subnational units such as states or provinces



## Box 2: What does standardization mean?

Tracking progress requires having tools capable of monitoring trends over time. For test scores to be meaningful over time, student performance must be measured against an inelastic yardstick of achievement. Tests must be standardized with respect to content and cognitive domains, format, administration procedures and scoring. Standardizing content and cognitive domains requires that the same or equivalent questions or performance tasks be posed for all students. Standardizing test administration requires uniformity in the written and verbal instructions given to students, in the length of time afforded them, in the materials provided to them, and in the physical testing environment. Standardized scoring requires explicit, impartial procedures for correcting tests or judging performance. It requires a standardized approach for accommodating test takers with special needs.

Source: Lockheed 1996

Monitoring learning progress (that is, positive change) requires measures that are both stable over time and standardized across time and location. Moreover, the equivalence of the measures must be confirmed (equated) empirically (Holland and Rubin 1982; Linn 2005). With respect to stability, assessments must be similar, if not identical, with respect to:

- content and cognitive domains and levels of difficulty
- reliability, validity and fairness
- target population definition and sampling strategy,
- how the assessments are administered,
- how the assessments are scored and results reported

Moreover, these similarities must be maintained over time, to enable the measurement of change. Table 2 summarizes these basic features of classroom assessments, examinations, national assessments and international assessments; only international assessments have the fundamental characteristics needed for cross-national monitoring.

Achieving a measurement instrument that meets the needs for cross-national monitoring has downsides, however. For example, the content and cognitive domains chosen for the assessment may exclude domains that are highly relevant in some countries, and the target populations as defined may exclude many children in some countries.

There are several reasons for excluding classroom assessments, examinations and national assessments **as they currently exist** as potential strategies for monitoring progress; Lockheed (2008) discusses these at length, and the next paragraphs summarizes the main points.

**Table 2: The who, what, why, when, where and how of assessments**

Assessment	Who? (target sample, population)	What? (content and cognitive domains)	Why? (purpose)	When? (periodicity)	Where? (assessment environment)	How? (technical and administrative procedures)
Classroom	Varies by teacher	Vary by teacher	Improve teaching	Any time	In classroom	Any method
Examinations	Varies by country, sub-national unit	Vary by country, sub-national unit	Selection Certification	End of cycle	In schools, testing centers, online	Standardized May be equated over time
National	Varies by country	Vary by country	Monitoring within country	Varies by country	In schools	Standardized May be equated over time
International	Fixed cross-country	Fixed cross-country	Monitoring cross-country	Systematic	In schools	Standardized Equated over time

*Classroom assessments*, with the diagnostic purpose of informing teachers about the progress of their individual students, are typically designed by individual teachers, conducted by individual teachers, scored by teachers and are not aggregated across teachers or classes. This means that the results of classroom assessments are rarely, if ever, available for purposes of broader monitoring.

*Examinations* may be highly standardized, but their purpose is to certify the accomplishment of individual students and/or to select individuals into the next higher level of education. This means they are poor choices for monitoring learning achievement, for several reasons. First, the level of difficulty of examinations can be set quite high, so that improvements at lower levels of performance are not registered. Second, comparability of scores over time and location can be compromised by selection effects;

students typically choose whether or not to sit for examinations and changes in the demographics of the student cohort can influence the average performance level. Third, in many countries “passing “ the examination can be influenced by the availability of spaces at the next level; this is often the case for examinations at the end of basic education or the end of upper secondary education.

*National assessments*, which also can be standardized within a given country, are designed to reflect the national curriculum and performance standards, which can vary across countries. Thus, at present, scores on national assessments are only minimally useful for cross-national monitoring and evaluation. As Benavot and Kosleki recently note: “national learning assessments are not designed for comparing learning outcomes across education systems” (2015:19). There are no “official exchange rates” or “purchasing power comparisons” that enable scores on national assessments to be compared across countries in the same way that national currencies can be compared.<sup>8</sup> Some national assessments use instruments that are equated with respect to domains and difficulty over time and therefore can provide valuable information about changes in learning outcomes within a single country. However, they do not provide useful information for cross-country comparisons since, for example, an improvement -- of 10% or half a standard deviation – in scores on one country’s national assessment could be equivalent to a completely different amount of improvement -- 2% or two standard deviations -- on the national assessment scores of another country, depending on the scale used<sup>9</sup>.

Could national assessments be equated, through various statistical techniques? Yes, in two ways. The first requires – at a minimum – tests that include a common set of questions (items) with similar statistical properties across countries. Numerous approaches to equating assessments have been tried over the past half century, with “discouraging results and cautions of many experts on equating” (Linn 2005: 20). The five requirements for equating, agreed upon by experts, are: (a) equal construct, (b) equal reliability, (c) symmetry, (d) equity, and (e) population invariance (Dorans and Holland 2000). If the national assessments of various countries cover the same content and cognitive domains, and sample the same student population, then equating might be possible, if the other requirements are also achieved. The content of virtually all national assessments include “language” and mathematics, suggesting that these two areas are universally relevant; other domains (science, social sciences, foreign languages) vary in importance across regions (Benevot and Koseleci 2015). The grade levels assessed in national assessment vary, but around 90% of countries with national assessment sample students in grades 4-6

---

<sup>8</sup> Eric Hanushek and Ludger Woessmann (2015) have used the US National Assessment of Education Progress to “equate” scores on various international assessments, but it is not clear that their methods meet professional standards for equating.

<sup>9</sup> National assessments tend to focus on primary grades 1-6, with some countries also assessing student learning outcomes at grades 8 or 9; very few countries implement national assessments at the end of upper secondary.

(Benevot and Koseleci 2015). These commonalities suggest that national assessments in “language” and mathematics could possibly be equated, but only if the tests themselves were designed with equating in mind. As Linn notes: “It is easy to see that the strict requirements of equating are unlikely to be met for assessments that are not specifically designed to be interchangeable” (Linn 2005: 21).

A second approach has been used to “equate” U. S. state assessment scores with international benchmarks, using a statistical linking methodology described in Johnson and others (2005). This process, called “chain linking” involves linking the state assessment scores to the U.S. national assessment (NAEP) scores, and then linking the NAEP scores to an international assessment (TIMSS or PIRLS) as described by Phillips (2014). This is possible because (a) students in grades 4 and 8 in all states are assessed with NAEP, and (b) in addition, students in the U.S. are assessed with PIRLS in grade 4 and with TIMSS in both grades 4 and 8. In other words, the requirements for equal construct (reading, mathematics or science proficiency) and population invariance (grade 4 or grade 8) are met; it can also be assumed that the assessments are equally reliable. To apply this approach in an international context would require that all countries participate in a common assessment such as TIMSS, PIRLS or PISA and that the national assessments in all countries include grades 4 and 8, or 15-year-olds, for example. At present these conditions are not met in most countries.<sup>10</sup>

*Conclusion: For monitoring progress toward the Goal 4 learning targets, classroom assessments and examinations are insufficiently standardized across countries to provide useful measures; national assessments could be used, but countries would need to (a) agree on the test design modifications required for this purpose, or (b) participate in an international assessment that could be utilized for equating purposes.*

---

<sup>10</sup> According to Benevot and Koseleci (2015): (a) fewer than 60 % of countries in Sub-Saharan Africa, Latin American and the Caribbean or the Arab states have conducted recent national assessment, and (b) among countries with recent national assessments, about 90% have assessed students in grades 4-6.

### 3 International large-scale assessments

Alternatives to classroom assessments, examinations and national assessments (as they currently stand) are international large-scale assessments (These are summarized in Annex A). Many highly standardized international and regional assessments for monitoring and evaluation now exist and some may be appropriate for monitoring change; these assessments cover selected levels of education and types of learners. Approximately two-thirds of all countries have participated or are currently participating in one or more of these assessments (Lockheed, Prokic-Breuer & Shadrova 2015).

The vast majority of international large-scale assessments target students at the primary level, 1<sup>st</sup>-6<sup>th</sup> grades (four assessments focus on grades 1-3 and seven focus on grades 4-6). Five assessments target students at the secondary level, 7<sup>th</sup> grade and above. Only one large-scale international assessment program – PIAAC – addresses youth and adults. No large-scale international assessment targets students in either post-secondary TVET or higher education, although PISA includes students enrolled in secondary-level vocational and technical programs. Two assessments, UWEZO and ASER<sup>11</sup>, include learners ages 5-6 years to 16 years, and are focused on foundational literacy and numeracy skills. The regional assessments focus on primary education only.

These international large-scale assessments (including regional assessments) are not fully aligned with the agreed-upon indicators for SDG Goal 4 (table 3). The only international large-scale assessments that could be used to assess readiness for primary school (target 4.2) – EGRA and EGMA – are administered to children who are already in school, omitting those who are not yet enrolled; they also do not measure other aspects of school readiness. The international large-scale assessments that could be used to assess learning outcomes at the end of the primary level (target 4.1) are targeted at two grades – grade 4 and grade 6 – and are thus less relevant to countries where primary school ends at grade 5. The international assessments that could be used to assess secondary level (also target 4.1) learning outcomes – PISA and TIMSS-- do not target students at the end of this level, but are reasonably appropriate for measuring learning outcomes at the end of lower secondary school. PISA measures the reading and mathematics performance of 15-year-olds, many of whom are studying in 9<sup>th</sup> grade, which is often the end of the lower secondary level, although some students are studying at 7<sup>th</sup> or 8<sup>th</sup> grades. TIMSS measures mathematics performance, but not reading, of 8<sup>th</sup> grade students. Only one international large-scale assessment -- TIMSS Advanced -- measures performance near the end of upper secondary school, but it assesses two science domains only. One international assessment – PIAAC – measures ICT skills of youth and adults (target 4.4). ASER and UWEZO measure foundational literacy and numeracy of youth to age 16, but not adults (target 4.6). PIAAC,

---

<sup>11</sup> ASER and UWEZO are often not included in lists of international large-scale assessments, and are listed here because they provide a measure of foundational literacy that has been used with out-of-school children and youth.

STEP and LAMP measure functional literacy of both youth and adults, and PIAAC and LAMP also measure functional numeracy (target 4.6).

**Table 3. How international large-scale assessments align with Goal 4 targets and indicators**

Goal 4 Target	Reading	Mathematics (	ICT
4.2 Readiness for primary	EGRA	EGMA	
4.1 End of primary	Grade 4: PIRLS Grade 6: ERCE, PASEC, SACMEQ	Grade 4: TIMSS Grade 6: ERCE, PASEC, SACMEQ	
4.1 End of lower secondary	Age 15: PISA	Grade 8: TIMSS Age 15: PISA PASEC	ICCS
4.1 End of upper secondary		Grade 11: TIMSS Advanced	
4.4 Youth/adults with ICT skills			PIAAC
4.6 Youth/adults with functional literacy and numeracy	ASER, UWEZO, PIAAC, LAMP	ASER, UWEZO, PIAAC, LAMP	

Note: ASER and UWEZO do not assess individuals over age 16.

*Conclusion: International large-scale assessments provide measures that are useful for monitoring some Goal 4 targets at the end of primary and lower secondary levels of education.*

## 4 Who, what, why, when, where and how of existing international large-scale assessments

The existing major large-scale international assessments are similar in only two respects: the “why” of the assessment, which is to measure learning achievement across countries, and the “where” of the assessment, which is typically “in school.” Otherwise, the assessments differ considerably with respect to who is the target population, what are content and cognitive domains of the assessment, and when and how the assessment is conducted.


*Who?* The SDG targets for monitoring learning achievement explicitly mention measurement “at the end of each level of education.” Countries vary greatly in their definition of how many years constitute each level of education, and the ISCED definitions for primary, lower secondary and upper secondary levels encompass many grade levels. Primary education (ISCED 1) ends after 4-7 years of schooling, but typically after 6 years. Lower secondary education (ISCED 2) ends after 8-11 years after the start of primary education, and upper secondary education (ISCED 3) ends after 12-13 years after the start of primary education. Students in many different grades, therefore, can be considered the target populations (the who) for assessment.

International large-scale assessments are highly variable with regard to whose learning they measure. Two assessments (ASER, UWEZO), target individuals 5 or 6 years of age to 16 years of age; one (PIAAC) targets adults ages 16-65; and one (PISA) targets 15-year-olds in school. Of the remaining 7 major large-scale international assessments – TIMSS (grades 4, 8, 11), PIRLS (grade 4), SACMEQ (grade 6)<sup>12</sup>, ERCE (grades 3, 6) and PASEC (grades 2, 6 and end of lower secondary) – only ERCE, PASEC and SACMEQ assess students in grade 6 and only TIMSS and PIRLS assess students in grade 4; all other assessments target students at different grade levels. Occasionally exceptions are made, as in the case of Botswana, Honduras and Yemen, where grade 6 students were tested using the TIMSS 2011 grade 4 instruments.

*What?* The assessments differ in what is assessed; that is, they do not measure the same content or cognitive domains. For example, although SACMEQ and ERCE both test students at grade 6, the tests themselves are different with respect to the content domains that are assessed. Most international large-scale assessments include assessments in the content domains of reading and mathematics, however. SACMEQ tests students in mathematics and reading, whereas ERCE tests students in reading, mathematics, writing and natural sciences. TIMSS assesses four content domains and three cognitive domains in both mathematics and science, but does not assess reading; PIRLS assesses reading but not other content domains. PISA assesses reading mathematics and science, with a complex framework of content and process domains. In addition, although reading and

---

<sup>12</sup> From 2014, PASEC is working with SACMEQ on the grade 6 assessment.



mathematics are typically among the domains assessed, performance levels are defined differently. This is discussed further in section 6.

*When?* The assessments differ with respect to when (and how often) the assessment is conducted. Assessments can differ with respect to the time in the school year that the assessment is administered, although generally they are administered in the second semester of the academic year. They can also differ with respect to the assessment's frequency. PISA is conducted every three years, TIMSS is conducted every four years, PIRLS is on a five-year cycle, and the regional assessments – ERCE and SACMEQ – have longer and irregular intervals for assessment.

*How?* The assessments differ with respect to how the assessment is undertaken. The how of assessments refers to all the various psychometric and operational tasks that go into creating, administering, scoring and reporting the results from an assessment. This is discussed in section 7.

*Conclusion: Different international assessments provide information about student learning outcomes at many different stages, but information post-primary is relatively limited.*



## 5 Main advantages and disadvantages of international and regional assessments

The advantages and disadvantages of these assessments depend somewhat upon the target ages and grades for the assessment (table 4).

*Primary grades 1-3.* For early grades, simple one-to-one oral assessments, carried out either in the child's home or at school under standardized conditions, may be preferable to written group assessments carried out in schools. These types of assessments, including ASER and UWEZO, provide information regarding foundational skills in reading and numeracy that are essential for a child to progress in primary school. The tests, themselves, however, are often very simple, involving few questions.<sup>13</sup> One-to-one assessments may be less threatening to some children than group assessments, although the one-to-one interaction with an unfamiliar adult may be threatening in some societies. The cost of these assessments can be quite low.

These assessments may not be entirely comparable across countries, however, particularly for foundational literacy, since written languages vary enormously linguistically. Main differences among languages include linguistic variations (agglutinative languages versus analytic languages) and orthographic variations (alphabetic languages versus ideogrammatic languages); in addition the degree of diglossia – the discrepancy between written and spoken versions of the same language – can affect performance. (Ferguson 1959). Thus, simple indicators derived from these assessments, such as “words per minute read” will have a different mean value for different languages. Comparability would need to be established by setting the results from the assessment on a common performance scale.<sup>1</sup>

In early grades, group administered tests that rely on unfamiliar testing formats – for example, multiple choice or short answers – may present challenges to some children. While group administration is less costly than one-to-one administration, children would need to be familiarized with the test formats in advance. Group standardized tests such as ERCE may be more reliable than individually administered tests, particularly those administered in home contexts, since the administration procedures can be observed by quality control monitors.

*Primary grades 4-6.* For the upper primary grades, group administered tests have been used for decades, apparently with little difficulty. These types of tests are appropriate when

---

<sup>13</sup> For example, the ASER test involves relatively few tasks. The test for reading requires the child to identify 10 letters of the alphabet, read 10 simple words, read a simple paragraph and read a short story; the mathematics test involves recognizing any five of the numbers 1-9, recognizing any five 2-digit numbers, solving two 2-digit subtraction problems with borrowing and solving one simple division problem.

students' reading comprehension is sufficient to respond to the written questions and problems. Guessing on tests using multiple choice formats, however, can result in scores that are no different than chance for many children in some countries. Many assessments are moving away from simple multiple-choice formats to include greater numbers of "constructed response" questions. These types of questions require professional scoring of answers, rather than machine scoring, to ensure standardization. TIMSS, PIRLS, ERCE, SACMEQ and PASEQ all use group administered tests with a combination of multiple choice and constructed response questions.

**Table 4: Advantages and disadvantages of existing international large-scale assessments**

Assessments	Advantages	Disadvantages
Oral assessments, grades 1-3 (example: EGRA, EGMA, ASER, UWEZO)	Explicit sampling procedures Simple administration Short duration Few psychometric challenges Low cost	No comparability across languages (linguistic variation, orthographic variation) No comparability across countries Requires many trained administrators One-to-one administration may disadvantage some children
Regional assessments, grades 1-3 (example: ERCE, PASEC)	Standardized cognitive instruments developed to avoid bias and have high reliability and validity Explicit sampling procedures Group administration following explicit instructions Scoring and reporting standardized Comparability across participating countries	Limited number of content domains Possible floor and ceiling effects in some countries
Regional and international assessments, grades 4-6 (example: ERCE, TIMSS, PIRLS, SACMEQ, PASEC)	Standardized instruments developed to avoid bias and have high reliability and validity Explicit sampling procedures Group administration following explicit instructions Scoring and reporting standardized Comparability across participating countries	Limited number of content domains Low discrimination at low end of scale Cost constraints for some countries May involve complex administration procedures (multiple test booklets, randomization of question blocks)
International assessments,	Standardized instruments developed to avoid bias and have high	Limited number of content domains

grades 7-12 (example: TIMSS, PISA)	reliability and validity Explicit sampling procedures Group administration following explicit instructions Scoring and reporting standardized Comparability across participating countries	Low discrimination at low end of scale Large scalar invariance for PISA reading assessment in middle-income countries (Asil & Brown 2016) Cost constraints for some countries May involve complex administration procedures (multiple test booklets, randomization of question blocks) Low coverage of age group in some countries
International assessments, ages 5-16 (example: ASER)	Explicit sampling procedures Simple administration Short duration Few psychometric challenges Low cost	No comparability across languages (linguistic variation, orthographic variation) No comparability across countries Requires many trained administrators One-to-one administration may disadvantage some children
International adult assessments (example: PIAAC, STEP, LAMP)	Standardized instruments developed to avoid bias and have reliability and validity Explicit sampling procedures Individual administration following explicit instructions Scoring and reporting standardized Comparability across participating countries	No assessments for end of higher education or specifically for vocational education

*Lower secondary grades 7-9.* At this level, only two assessments -- TIMSS for Grade 8 and PISA for students age 15 that are enrolled in Grade 8 or 9 -- measure learning achievement near the end of this level of education. Both assessments use group administered tests in school, with TIMSS using a grade-based sample and PISA an age-based sample. That is, all students in the TIMSS sample have completed grade 7 and are currently in grade 8. By comparison, students sampled for PISA can be studying in any grade from grade 7 onward. This leads to substantial differences across countries with respect to what content students will have had the opportunity to study. For example, for PISA 2012, more than

80% of students were enrolled in grade 9 or below in 14 countries, while, more than 80 percent of students were enrolled in grade 10 or above in 21 other countries (OECD 2014). *Upper secondary, grades 10-12.* At the upper secondary level, few assessments are available. Only TIMSS Advanced and PISA provide any information regarding student learning achievement at this level. Both have shortcomings for monitoring learning achievement related to the SDG targets. TIMSS Advanced covers only the content domains of mathematics and physics, and does not assess reading achievement. PISA assesses both reading and mathematics, but its target population is 15-year-olds, generally enrolled in grades 9 or 10 (considered the “modal grades” in most countries). Students studying at the end of upper secondary, grades 11 or 12, were considered the “modal grade” for only two countries participating in PISA 2012 (New Zealand and the UK). In these two countries, 80% or more of 15-year-olds were enrolled in grades 11 or 12.

*TVET, higher education and adult skills.* In general, no international assessments measure learning outcomes for TVET or higher education. Within household surveys, two international assessments – STEP and LAMP – measure adult literacy and LAMP also measures numeracy. One international assessment, PIAAC, measures adult skills of literacy, numeracy, reading and problem-solving in a technology rich environment.

*Conclusion: Existing international large-scale assessments can provide useful information for monitoring learning achievement in reading and mathematics at the end of primary and lower secondary education levels. However, three challenges to using these assessments remain: (a) linguistic issues may limit the cross-national comparability of existing assessments of early reading; (b) reading scales may not be equivalent across countries; and (c) the target populations of existing assessments may limit their application for monitoring learning achievement at the end of secondary school. The lack of any assessments for measuring post-secondary learning outcomes is a major challenge to monitoring the SDG targets at these levels.*

## 6 What is monitored: Comparison of performance standards

The major international large-scale assessments provide not only average scores for student performance in the content and cognitive domains that are assessed, but also “performance levels” that range from low to high. Performance levels seek to identify the skills needed to carry out certain types of activities. Comparison of performance standards across the different international large-scale assessments is difficult, due to the differences in target populations and therefore often differences in standards for performance.

In some cases, the performance standards appear to be similar. This would be the case for comparing PIRLS reading standards (for grade 4 students) with PISA reading standards (for 15-year olds). For example, the PIRLS “basic” reading level (low international benchmark) requires students to “locate and retrieve an explicitly stated detail” [in a literary text] or “locate and reproduce explicitly stated information that is at the beginning of the text” [in informational texts] (reference), and the PISA basic reading level (Level 1b) requires the students to “locate a single piece of explicitly stated information in a prominent position in a short, syntactically simple text with a familiar context and text type, such as a narrative or a simple list” (OECD 2014: 191). These appear quite similar, but – because of differences in the populations sampled – the share of students reaching the PIRLS basic level in reading is much lower than the share of students reaching the PISA basic level in reading, in any country. For example, 66% of 4<sup>th</sup> grade students compared with 96% % of 15-year-olds in Indonesia reached the basic level as defined in these two assessments.

In other cases, standards for closely related grades can call for different skills. For example, the basic level for TIMSS mathematics (for grade 4) and for TERCE mathematics (for grade 3) can be compared. The TIMSS “basic” mathematics level is defined as: “Students have some basic mathematical knowledge. Students can add and subtract whole numbers. They have some recognition of parallel and perpendicular lines, familiar geometric shapes, and coordinate maps. They can read and complete simple bar graphs and tables.”

By comparison, the TERCE “basic” mathematics level is defined as: “Students can recognize the relationship of order between natural numbers and common two-dimensional geometric figures in simple drawings. They can locate relative positions of an object in a spatial representation. They can interpret tables and graphs in order to extract direct information.” The standards for TERCE appear easier than those for TIMSS, and the share of students who reach the basic level is higher in TERCE than in TIMSS. For example, 95% of TERCE grade 3 students compared with 77% of grade 4 students in Chile reached the basic level as defined in these two assessments.

The full performance standards in mathematics as described by TIMSS (for grade 4) and TERCE (for grade 3) can be found in Annex B

One way to compare the performance levels established for two different assessments is to observe the distribution of student performance in a given country across the levels as defined by the two assessments. The assumption is that – if the performance levels were equivalently defined – student performance in that country would be similarly distributed across the levels on both assessments. From the existing international large-scale assessments, there are very few opportunities to make such comparisons.

Chile and Botswana, however, provide two opportunities. Chile participated in both TIMSS 2011 and TERCE (in 2013), and Botswana participated in both SACMEQ III (in 2007) and TIMSS 2011. In both countries, the performance distribution of students differs between the two assessments (tables 5 and 6).<sup>14</sup>

In Chile, the differences are obvious. A much higher share of students fail to reach the lowest performance level on TIMSS, as compared with TERCE. And TERCE reports a much higher share of students reaching the highest performance level, as compared with TIMSS. These differences cannot be explained by the differences in the grade levels (at the low end, more 4<sup>th</sup> grade students should reach at least a basic level of performance as compared with 3<sup>rd</sup> grade students, and at the high end, students don't improve that rapidly between one grade and the next). The observed differences, therefore, can be attributed to differences in the definitions and measurements of the skills levels. This has implications for making comparisons using the two assessments.

**Table 5: Mathematics performance levels, Chile, Grade 3 (TERCE 2013) and Grade 4 (TIMSS 2011)**

	Mathematics performance				
	< Level 1	Level 1	Level 2	Level 3	Level 4
% at each level					
TIMSS (4 <sup>th</sup> )	23%	17%	28%	12%	2%
TERCE (3 <sup>rd</sup> )	5%	28%	34%	19%	14%
% reaching each level					
TIMSS (4 <sup>th</sup> )	23%	77%	44%	14%	2%
TERCE (3 <sup>rd</sup> )	5%	95%	67%	33%	14%

The story in Botswana is more complicated, because SACMEQ uses eight performance levels for mathematics, compared with the five used by TIMSS. In addition, SACMEQ III

<sup>14</sup> Moreover, in their reports, TIMSS and ERCE report performance levels differently. TIMSS reports the percentage of students reaching each level, so the percentages reported for each lower level are cumulative (that is, students who reached Level 2 also reached level 1, so the Level 1 percentage includes those who reached each subsequent level). ERCE reports the percentage of students performing at their maximum level, so the percentages are not cumulative. Table x presents both indicators.

was administered four years before TIMSS 2011, so some changes in teaching could have occurred. The eight SACMEQ performance levels for mathematics begin with “pre-numeracy” and end with “abstract problem solving.” These can roughly be mapped to the TIMSS performance standards, as in table 6.<sup>15</sup>

**Table 6: Mathematics performance levels, Botswana, Grade 6, SACMEQ 2007 and TIMSS 2011**

		Mathematics performance				
		< Level 1 (SACMEQ 1 & 2)	Level 1 (SACMEQ 3)	Level 2 (SACMEQ 4 & 5)	Level 3 (SACMEQ 6 & 7)	Level 4 (SACMEQ 8)
% at each level						
TIMSS		40%	31%	22%	7%	0%
SACMEQ		22%	34%	36%	7%	0.4%
% reaching each level						
TIMSS		40%	60%	29%	7%	0%
SACMEQ		22%	78%	44%	7.4%	0.4%

Again, the share of students in Botswana who performed at various levels of mathematics differs between TIMSS and SACMEQ. On the TIMSS performance scale, which was set for 4<sup>th</sup> graders internationally, 40 percent of Botswana students failed to reach minimum competency, whereas on the SACMEQ scale, only 22% of students failed to reach this level of competency. Similarly, a much higher share of students performed at the basic level on the SACMEQ assessment (level 3) as compared with the TIMSS assessment (level 1), as well as on the “intermediate” level (level 2 TIMSS, SACMEQ levels 4 & 5). On both assessments, approximately the same share performed at “high” or “advanced” levels. Although the SACMEQ scores could be “cut” at different points, to achieve greater comparability with TIMSS, the performance standards for TIMSS appear to be set somewhat higher than those on SACMEQ.

The point of these comparisons is not to critique the assessments, but rather to demonstrate the difficulty of making simple comparisons across them. Differences in how performance levels are defined, how instruments are constructed and what content and cognitive domains (including difficulty levels) that each assessment covers would need advanced psychometrics --empirical equating, IRT scaling -- prior to using multiple assessments for comparison across countries.

*Conclusion: Scale scores and performance standards in international large-scale assessments differ across assessments of the same content and cognitive domains.*

<sup>15</sup> As a point of comparison, the average score for Botswana students on the 4<sup>th</sup> grade TIMSS 2011 was 419 (s.e.= 3.7), compared with the average score on the 6<sup>th</sup> grade SACMEQ III, which was 535 (s.e = 4.5).

## 7 How students are assessed

International large-scale assessments entail a great number of discrete activities as well as the coordination of these activities across countries. These can be roughly grouped into technical activities and administrative activities.

*Technical activities.* The technical activities related to international large-scale assessments are similar to the technical activities related to any large-scale assessment intended to monitor trends over time. These involve test development, through classical or modern techniques, of assessment tools that are reliable, valid, and fair. Most assessments sample skills and abilities that are not directly observable, and must be inferred from performance on the assessment. Analyses of reliability, validity and fairness ensure that these inferences are appropriate.<sup>16</sup>

Reliability refers to the ability of an assessment tool to measure skills and abilities consistently. This is essential for monitoring performance over time. If an assessment tool uses an “elastic meter-stick”, then any inferences about changes in performance will be misleading. Reliability is a pre-requisite for validity.

Validity is a matter of degree, and no assessment tool is absolutely valid or absolutely invalid. Validity research and analyses establish that the assessment tool (or “test”) measures what it is supposed to measure. In particular, the assessment tool should measure both the skills and abilities that should be measured (construct validity) and the appropriate content (content validity); it should also be able to predict “success” as appropriate (predictive validity); it should entail few adverse consequences (consequential validity); and it should have an expected relationship with other measures of the same construct (external validity). Establishing both construct and content validity is an essential activity of test development.

Fairness in assessment is important for ensuring a level playing field for those taking the test. Assessment experts conduct several different analyses designed to identify bias in construct validity, content validity and predictive validity. These analyses include fairness evaluations by trained reviewers and routine analyses of test questions to determine whether or not particular questions unfairly contribute to group differences (“differential item functioning” or DIF analyses). Fairness training for coders of open-ended responses and accommodation for those with disabilities or with health-related needs also ensure fairness.

International large-scale assessments are developed with these considerations of reliability, validity and fairness.

---

<sup>16</sup> The following three paragraphs are drawn from materials of the College Board and Educational Testing Service.



Scientific sampling of students is also an important aspect of international large-scale assessments. These assessments typically employ multi-stage sample (for example: schools, classrooms, students) and have standardized approaches for dealing with replacement samples. For both IEA's and OECD's assessments, sampling is carried out with the use of specialized sampling software.


*Administrative activities.* Once the content and cognitive domains for the assessment have been determined, and the target population identified, the following tasks must be completed (this is not an exhaustive list, and many of these tasks will be undertaken twice, once for the piloting of the assessment and once for the actual assessment):

- item writing and review
- test assembly
- translation of tests and survey instruments
- gaining permission to approach schools
- sampling the target population
- printing or layout of test booklets (or working with platform for computer-based assessment)
- packaging and distributing materials to schools
- administering assessment in schools
- training of coders
- coding of constructed-response test and questionnaire items
- data entry and quality control
- submitting data to international coordinating teams
- reviewing feedback from international experts
- reviewing cleaned data set
- scale score construction or interpretation
- equating and trend analysis
- report writing

In the case of international large-scale assessments, some of these tasks are completed by technical experts. In particular, once items have been written, translated and pilot tested in all countries<sup>17</sup>, the technical experts are able to assemble one or more test booklets (or prepare one or more computer-based or computer adaptive tests) that need only to be duplicated in the participating countries. Similarly, once the target population has been identified, countries may use sampling software produced by technical experts to actually draw the relevant sample of students for the assessment from lists (typically of schools, classes and students) at the national level.

---

<sup>17</sup> Pilot testing of items will establish various item parameters, including item difficulty, distractor functioning, and such indicators of bias as differential item functioning (DIF) for population subgroups.



Despite this assistance, countries participating in international large-scale assessments report challenges for many of these tasks, including item writing, translation, sampling, access to schools (particularly in federal-type countries), computer-based assessment needs, survey administration, coder training, coding, data submission and preparing national reports (Lockheed, Prokic-Breuer & Shadrova 2015). These challenges indicate a lack of assessment capacity in selected areas.

Most sponsors of international large scale assessments build capacity by offering training for many of these tasks. In addition, international donors, universities, professional associations and assessment institutions have provided short-term training relevant to these tasks. Publications detailing the numerous steps of undertaking national and international assessments also are available. And international donors have provided support through direct support for training and via support to countries for participation in such assessments. Learning-through-doing is a important strategy for building capacity.

*Conclusion: International large-scale assessment are complex undertakings, and considerable support is needed for countries that lack strong assessment capacity.*

## 8 Why countries do not already participate in international large-scale assessments

Countries may not participate in international large-scale assessments for any number of reasons, which can be summarized as the challenges of culture, capacity and costs (Bloem 2013, Lockheed 2010). Challenges of capacity and costs may be addressed through donor support; challenges to participation related to contextual differences among countries may be difficult to overcome.

*Culture.* There is little doubt that interest in international large-scale assessments has grown sharply over the past decade. Approximately two-thirds of all countries with populations greater than 30,000 have participated in one or more international or regional large-scale assessment (Lockheed, Prokic-Breuer & Shadrova 2015). Interest in participating in international large-scale assessments could grow substantially if indicators based on such assessments were adopted for monitoring the SDGs. But many low and lower-middle income countries do not participate in such assessments. A recent analysis of middle-income countries participating in PISA notes that participation rates for high-income and upper-middle-income countries are many times higher than participation rates for lower-middle-income and low-income countries (Lockheed 2015).

Culture may play a role in why these countries avoid participating in international assessments. Most international large-scale assessments have involved countries in the developed world, which share many similar cultural values (Hofstede, Hofstede and Minkov 2010). These cultural values may differ in other countries, and may influence how national results on international assessments are perceived by government officials. In particular, while unexpectedly poor results on international large-scale assessments have often stimulated a call for action and education reform in OECD countries, poor results are considered “shameful” or “embarrassing” in countries with a different cultural context. For example, when Mexico scored below expectations on TIMSS, it declined to publicly release its scores, and in Georgia, PISA 2012 was cancelled after the country performed poorly on both TIMSS and PISA.

*Assessment capacity.* The capacity for undertaking the activities needed for participating in an international large-scale assessment can vary enormously across countries. In general, high income countries tend to have well-developed assessment systems with the capacity to participate in international large-scale assessments, while low- income and lower middle-income countries do not (Lockheed, Prokic-Breuer & Shadrova 2015). In the relatively few low- and lower middle-income countries whose capacity for undertaking international large-scale assessments has been directly measured, the conclusions are disturbing: most are judged to have “latent” (that is, none) or “emerging” (that is, only partial) capacity for undertaking these assessments (World Bank, SABER Student assessments).<sup>18</sup> The capacity

---

<sup>18</sup> SABER-Student Assessment website and country reports [www.saber.worldbank.org/index.cfm](http://www.saber.worldbank.org/index.cfm)

for undertaking international large-scale assessments appears to be higher in countries with long-standing national assessment systems. These countries are more likely to participate in an international assessment such as PISA (Lockheed 2015).

*Costs.* Cost considerations appear to be a constraint for participation in an international large-scale assessment. The direct costs of participation in an international large scale assessment are very small—less than one-tenth of one-percent -- when expressed as a share of a country's total expenditure on education (Wagner, Babson & Murphy 2011; Wolff 2007). While the costs of participating in an international large-scale assessment are relatively modest, these costs – particularly foreign exchange -- can serve as barriers to participation. Costs that are typically mentioned include the international participation fees (45,500 EUR per year over four years for PISA and \$20,00 per year over five years for PIRLS or TIMSS, for example), travel and subsistence for attending mandatory international training meetings related to the various stages of the assessment, and extra-budgetary expenditures associated with the variable costs of the assessment, such as hiring temporary staff, purchasing materials to produce the assessments (including renting or purchasing computers for computer-based assessments), and local travel.

Several countries have received support from international donors to enable their participation in international large-scale assessments. For example, about one-third of education projects approved by the World Bank from 1998-2009 supported one or more international large-scale assessment: 8 projects supported PIRLS, 19 projects supported PISA and 21 projects supported TIMSS. Among 8 more recent World Bank education projects, 6 supported PISA 2 supported TERCE, 4 supported TIMSS and 1 supported PIRLS; several projects supported more than one assessment (Lockheed, Prokic-Breuer & Shadrova 2015). The most frequent support is for the payment of the international participation fees. For example, such donors as the World Bank and the UNDP assisted many low and middle-income countries with international participation fees and other expenses for several cycles of TIMSS and PIRLS: 18 countries for TIMSS 1999, 29 countries for PIRLS 2001 and/or TIMSS 2003, and 20 countries for PIRLS 2005 and/or TIMSS 2007 (Lockheed 2008; Lockheed, Prokic-Breuer & Shadrova 2015).

*Conclusion: Countries that do not already participate in international large-scale assessments face challenges of insufficient assessment capacity, the perception that participation involves high costs and cultural constraints. Technical and financial support has enabled many low and middle-income countries to participate in various international large-scale assessments, but many do not.*

## 9 Use of international large-scale assessments

Sponsors of international large-scale assessments typically emphasize the use of these assessments in education policy agenda setting and reform. The evidence for these claims is, however, extremely limited, particularly with regard to effects in low and middle income countries. In a few cases, countries have declined to reveal or include their country's results in international reports and to release their data sets for analysis; this may limit the assessment's utility to some extent (confidential within-country analyses may be undertaken, however).<sup>19</sup>

*Agenda-setting.* With respect to agenda-setting, international large-scale assessments have documented issues of education quality and equity, particularly in the upper-middle income and high income countries that participate in such assessments. The effect on agenda-setting in low and lower-middle income countries is less pronounced. For example, a systematic review of the literature from 1990-2011 identified 19 studies (11 of which were rated as "high quality") that addressed the impact of international assessments on education policies in low and middle-income countries (Best et al., 2013a). The studies that were reviewed included seven international large-scale assessments: TIMSS, PIRLS, PISA, International Assessment of Education Progress (IAEP), IEA Civic Education Study (CIVED), International Computer Competence Study (ICCS), and Monitoring Learning Achievements (MLA). The review found that international large-scale assessments were less associated with agenda setting or policy formation than with policy implementation and monitoring and evaluation. By comparison, a recent survey of 6,744 "opinion leaders" in 126 low- and middle-income countries and jurisdictions, examined the policy-making influence of external assessments across multiple sectors, including education. The study concluded that external assessments were more influential on agenda setting than on specific policy design (Parks et al., 2015).<sup>20</sup>

*Specific education policies and practices.* In a few cases, the effects of large-scale international assessments on specific education policies and practices have been documented. In general, however, the impact of international large-scale assessments has not been rigorously evaluated. Evaluations of TIMSS and PISA note that the results of these

---

<sup>19</sup> Explicit examples are difficult to identify, since a country's absence from an international report can also reflect technical shortcomings in survey implementation.

<sup>20</sup> PISA was the only international large-scale assessment that was included in the list of potential influencers, which included: the UNESCO Global Monitoring Report, the World Bank's Education Sector Review, the World Bank's EdStats and the Paris Declaration indicators. The exclusion of the other major international large-scale assessments (TIMSS, PIRLS, SACMEQ, TERCE) from the list raises questions about how comprehensive the study was with respect to education. Only 5% of the respondents (377) self-identified as working in the education sector, and only 15 respondents rated PISA's influence. The survey included respondents from internationally unrecognized jurisdictions, such as "Kurdistan" and "Puntland".

assessments have informed reforms in education performance standards, assessments, curricula and instructional materials and teacher professional development (Gilmore 2005, Elley 2002, Lockheed 2008, Breakspear 2012). Parks et al. (2015), however, found that respondents perceived that actual education reforms were only weakly influenced by external assessments.

The perceptions of stakeholders from middle-income countries regarding the effects of international large-scale assessments on education policy differ from the perceptions of stakeholders from high-income countries. For example, an OECD survey of stakeholders from countries participating in the first three PISA cycles reported that PISA positively affected education policy with respect to: development of national standards, establishment of national institutes of evaluation, changes in the curriculum, introduction of targeted educational programmes, increased allocation of resources to schools, and increased collaboration among key stakeholders (OECD, 2008). However, among the 12 middle-income countries that responded to this survey, 7 reported that PISA had “relatively low levels of impact on policy formation”, 2 reported “relatively medium levels of impact” and only 3 (Mexico, the Kyrgyz Republic and Thailand) reported “relatively high levels of impact” (OECD 2008; Breakspear 2012).

Analyses of the results from large-scale international assessments have suggested five broad areas where changes in education policy could result in higher student learning outcomes in low- and middle-income countries: 1) selecting and grouping students; 2) non-personnel resources invested in education; 3) resources invested in the quality of instructional staff; 4) school governance and assessments; and 5) curriculum and instruction (Glewwe et al., 2014; Kremer, Brannen, & Glennerster, 2013; Krishnaratne et al., 2013; McEwan, 2014; Murnane & Ganimian, 2014; OECD, 2013).

Studies focused on specific middle-income countries indicate that international large-scale assessments typically affect curriculum standards, performance targets and – in some cases – specific education reform policies intended to boost performance. For example, Jordan responded to both TIMSS and PISA results to compare itself with the world’s best achievers, review its curriculum, establish performance benchmarks and revise teacher training (Abdul-Hamid et al., 2011). In the Kyrgyz Republic, PISA results affected reforms such as the development of new standards and curricula, reductions in teaching load, upgrading of physical facilities, teaching practices and per-capita financing (Shamatov, 2014; Shamatov & Sainazarov, 2010). Indonesia and Turkey adopted performance targets and indicators based on PISA (Breakspear 2012).

## 10 Conclusions and lessons learned

### *Conclusions*

This paper has reached a number of conclusions, as follows:

- For monitoring progress toward the Goal 4 learning targets, classroom assessments and examinations are insufficiently standardized across countries to provide useful measures.
- National assessments, which may be highly standardized within any country, would need to be adapted to provide useful measures for monitoring Goal 4 learning targets.
- International large-scale assessments currently provide measures that are useful for monitoring Goal 4 targets at the end of primary and lower secondary levels of education, but provide little information about student learning outcomes at the end of upper secondary or post-secondary levels.
- Existing international large-scale assessments can provide useful information for monitoring learning achievement in reading and mathematics at the end of primary and lower secondary education levels. Linguistic issues may limit the cross-national comparability of existing assessments of early reading.
- Scale scores and performance standards in international large-scale assessments differ across assessments of the same content and cognitive domains.
- International large-scale assessments are complex undertakings, technically and administratively.
- Countries that do not already participate in international large-scale assessments face challenges of insufficient assessment capacity, the perception that participation involves high costs, and cultural constraints. Technical and financial support has enabled many low and middle-income countries to participate in various international large-scale assessments.

### *Lessons learned*

*“To everything there is a season, and a time to every purpose under the heaven.”* This applies as much to the assessment of student learning outcomes as to anything else. For the purpose of global education monitoring, the time is for international large-scale assessments. But for other purposes, other assessments have their time as well:

- Classroom assessments are valuable in helping teachers to identify the learning needs of their students and to adjust their instructional practices to these needs.
- As education systems continue to expand opportunities for students at all levels, selection into higher levels will become less important. The purpose of examinations will transition from selection to certification. Many professions require certification examinations, designed to guarantee that holders of the

certificate have the necessary skills to do their jobs, whether they are plumbers of physicians.

- National assessments are essential for countries to monitor their learning of their own students, according to their own national standards.<sup>21</sup>

International large-scale assessments are necessary for global monitoring. The existing ones have three main shortcomings, however:

- They do not completely align with the targets of Goal 4, with respect to the levels of education and the content domains needed for monitoring Goal 4 targets.
- Existing tests of early literacy, which have been applied in many countries, may be relevant only for alphabetic languages, particularly those that have little orthographic variation (unlike English, for example). Ideogrammatic languages, which use characters or symbols to represent concepts, require more memorization than alphabetic languages, and hence measures of fluency may be quite different.
- Existing international-large scale assessments, particularly PISA and TIMSS, were developed primarily for upper-middle-income and high-income countries. These assessments may require some modifications to improve their utility for low-income and lower-middle-income countries.
- Existing international large-scale assessments tend to focus on countries in different world regions, and their measures can be linked empirically for only a few, typically high-income, countries.

In response to the third shortcoming, both OECD and IEA are in the process of adapting their principal assessment tools for use in a wider range of countries. OECD is undertaking “PISA for Development” in six low and lower-middle-income countries (OECD 2015). IEA is developing “Literacy and Numeracy Assessment: (LaNA) for 4<sup>th</sup>-6<sup>th</sup> grade students in low and lower-middle-income countries, which will be linked with TIMSS and PIRLS (IEA 2015).


In response to the fourth shortcoming, the following are recommendations for research and studies that could improve the utility of existing large-scale international assessments for monitoring global learning outcomes:

- Equating studies (that is, incorporating a “block” of identical test questions into two different assessments, or using computer adaptive testing), which could enable performance on one assessment to be “horizontally” linked with performance on another assessment; some possible opportunities for equating studies:
  - Between SACMEQ/PASEC and ERCE
  - Between PIRLS and ERCE

---

<sup>21</sup> But, as studies in the United States have shown, not all jurisdictions adhere to the same standards, and what is “high performance” in any given jurisdiction may be considerably lower in a different jurisdiction (Phillips 2014).





Between TIMSS (4<sup>th</sup> grade) and ERCE  
Between PISA and TIMSS (8<sup>th</sup> grade)

- Equating studies among national assessments for literacy and numeracy at key grades
- Predictive validity studies for EGRA and EGMA in different languages
- Fairness/bias studies for national assessments and regional assessments that do not currently conduct such studies
- Studies and analyses that would place multiple assessments on a similar vertical scale

## References

- Abdul-Hamid, H., K. Abu-Lebdeh and H. Patrinos (2011), "Assessment Testing Can Be Used to Inform Policy Decisions", *Working Paper Series*, WPS5849, World Bank, Washington, DC.
- Asil, M. & G. T. L. Brown (2016), "Comparing OECD PISA Reading in English to other languages: Identifying potential sources of non-invariance", *International Journal of Testing*, 16:1, 71-93, DOI: 10.1080/15305058.2015.1064431
- Benavot, A and N. Koseleci (2015) "Seeking quality in education: The growth of national learning assessments, 1990-2013". Paper commissioned for the EFA Global Monitoring Report 2015, Education for All 2000-2015: Achievements and Challenges.
- Best, M., P. Knight, P. Lietz, C. Lockwood, D. Nugroho and M. Tobin (2013), The impact of national and international assessment programmes on education policy, particularly policies regarding resource allocation and teaching and learning practices in developing *countries*, final report, EPPI-Centre, Social Science Research Unit, Institute of Education, University of London, London.
- Breakspeare, S. (2012), "The Policy Impact of PISA: An Exploration of the Normative Effects of International Benchmarking in School System Performance", OECD Education Working Papers, No. 71, OECD Publishing, Paris.  
<http://dx.doi.org/10.1787/5k9fdqffr28-en>.
- Clarke, M. (2012), *What matters most for student assessment systems: A framework paper*, The World Bank, Washington, DC.
- Cresswell, J., U. Schwantner and C. Waters (2015). *A Review of International Large-Scale Assessments in Education: Assessing Component Skills and Collecting Contextual Data*, PISA, The World Bank, Washington, CD, OECD Publishing Paris.
- Elley, W. (2002), *Evaluating the impact of TIMSS-R (1999) in low-and middle income countries: An independent report on the value of World Bank support for an international survey of achievement in mathematics and science*, International Association for the Evaluation of Educational Achievement (IEA), Amsterdam.
- Ferguson, C. A. (1959), *Principles of teaching languages with diglossia*. Monograph Series on Languages and Linguistics, 437.
- Gilmore, A. (2005), *The impact of PIRLS (2001) and TIMSS (2003) in low-and middle-income countries*, International Association for the Evaluation of Educational Achievement (IEA), Amsterdam.

- Glewwe, P., E. Hanushek, S. Humpage, and R. Ravia (2014), "School resources and educational outcomes in developing countries: A review of the literature from 1990 to 2010", in P. Glewwe (ed.), *Educational Policy in Developing Countries*, University of Chicago Press, Chicago.
- Hanushek, E. A., & Woessmann, L. (2015), *The knowledge capital of nations: Education and the economics of growth*. MIT Press.
- Holland, P., & Rubin, D. (1982), *Test Equating*. Academic Press. New York.
- Johnson, E. G., Cohen, J., Chen, W., Jiang, T., & Zhang, Y. (2005). *2000 NAEP–1999 TIMSS linking report*. Washington, DC: National Center for Education Statistics, U.S. Department of Education.
- Kremer, M., C. Brannen and R. Glennerster (2013), "The challenge of education and learning in the developing world," *Science*, Vol. 340/6130, pp. 297-300.
- Krishnaratne, S., H. White and E. Carpenter (2013), *Quality education for all children? What works in education in developing countries*, Working Paper 20, International Initiative for Impact Evaluation, New Delhi.
- Linn, R. L. (2005). *Issues in the Design of Accountability Systems*. CSE Technical Report 650. National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Lockheed, M. E. (1996), "Assessment and management: World Bank support for educational testing", in A. Little & A. Wolf (eds.) *Assessment in Transition: learning, monitoring and selection in international perspective*, Oxford: Elsevier
- Lockheed, M. (2008), *Measuring Progress with Tests of Learning: Pros and Cons for "Cash on Delivery Aid" in Education*. Center for Global Development Working Paper, (147)
- Lockheed, M. (2010), *The Craft of Education Assessment: does participating in international and regional assessments build assessment capacity in developing countries?*, International Association for the Evaluation of Educational Achievement (IEA), Amsterdam.
- Lockheed, M., T. Prokic-Bruer and A. Shadrova (2015), *The Experience of Middle-Income Countries Participating in PISA 2000-2015*, World Bank, Washington, D.C./OECD Publishing, Paris. <http://dx.doi.org/10.1787/9789264246195-en>
- McEwan, P. (2014), "Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments", *Review of Educational Research*,
- Murnane, R. and A. Ganimian (2014), *Improving educational outcomes in developing countries: Lessons from rigorous evaluations*, National Bureau of Economic Research.

Murphy, P., V. Greaney, M. Lockheed and C. Rohas (eds.) (1996), *National Assessments: Testing the System*, The World Bank, Washington DC.

OECD (2013), *PISA 2012 Results: What Makes Schools Successful (Volume IV): Resources, Policies and Practices*, PISA, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264201156-en>.

OECD (2014), *PISA 2012 Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science (Volume I, Revised edition, February 2014)*, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264208780-en>.

Parks, B., Z. Rice and S. Custer (2015), *Marketplace of Ideas for Policy Change: Who do Developing World Leaders Listen to and Why?* AidData and The College of William and Mary, Williamsburg, VA.

Phillips, G. (2014), *International benchmarking: State and national education performance standards*. Washington, D.C.: American Institutes for Research

Ross, K. N., & Mählck, L. (Eds.). (1990). *Planning the quality of education: the collection and use of data for informed decision-making*. Paris: Unesco.

Shamatov, D. (2014), "Education Quality in Kyrgyzstan and the Programme for International Student Assessment (PISA)", in *Qualities of Education in a Globalised World*, pp. 43-62, Springer.

Shamatov, D. and K. Sainazarov (2010), "The impact of standardized testing on education quality in Kyrgyzstan: The case of the Program for International Student Assessment (PISA) 2006", in A.W. Wiseman (ed.), *The Impact of International Achievement Studies on National Education Policymaking*, Vol. 13, Emerald Group Publishing Limited.

Wagner, D., A. Babson and K. Murphy (2011), "How Much Is Learning Measurement Worth? Assessment Costs in Low-Income Countries", *Current Issues in Comparative Education*, Vol. 14(1), pp. 3-21. Teachers College, Columbia University, New York.

Wolff, L. (2007), *The costs of student assessments in Latin America*, PREAL, Washington DC.

World Bank SABER

## Annex A: Features of Existing International Large-Scale Assessments

Education level	Assessment	Age/grade	Countries	Content domains	Periodicity	Administration	Results
Preschool	ECES	End of preschool	tbd				
All school age	ASER	Age 5-16	Rural India	Reading Math English	Annual	One-to-one oral in homes	Performance levels
	UWEZO	Age 6-16	Kenya, Tanzania, Uganda	Basic literacy and numeracy	Annual	One-to-one oral in homes	Scores
Primary (grades 1-3)	EGMA	Grade 1-3	11	Numeracy mathematics	On demand	One-to-one oral in school	Scores
	EGRA	Grade 1-3	42 (est)	Literacy (foundation skills)	On demand	One-to one oral in school	Scores
	PASEC	Grade 2	10	Reading, writing, numeracy	On demand	Group administration in school	Scores
	ERCE	Grade 3	15	Math Reading Writing	5-6 years	Group administration in school	Scores
Primary (grades 4-6)	PIRLS	Grade 4	50	Reading	5 years	Group administration in school	Scores Performance levels
	TIMSS	Grade 4	59	Math Science	4 years	Group administration in school	Scores Performance levels
	PIRLS Literacy	Grade 4, 5, 6	Botswana, Colombia, South Africa	Literacy	5 years	Group administration in school	Scores Performance levels
	ERCE	Grade 6	15	Math Reading Writing Natural sciences	Not fixed	Group administration in school	Scores
	LaNA	End primary	tbd	Literacy Numeracy	4 years	Group administration by in school	Scores

## Annex A: Features of Existing International Large-Scale Assessments

Educational level	Assessment	Age/grade	Countries	Content domains	Periodicity	Administration	Results
	PASEC	Grade 6	10	Reading, writing, numeracy	First international in 2014	Group administration by in school	Scores
	SACMEQ	Grade 6	15	Reading and mathematics	Not fixed	Group administration in school	Scores
Lower Secondary	PISA	Age 15 in grades 7-9	22 (9 <sup>th</sup> grade is modal grade for 2012)	Math Reading Science	3 years	Group administration in school; optional computer-based administration	Scores Performance levels
	TIMSS	Grade 8	59	Math Science	4 years	Group administration in school; optional computer-based administration	Scores Performance levels
	ICILS	Grade 8	21	Computer literacy Information literacy	Once (2013)	Computer-based administration	Scores Performance levels
Upper Secondary	PISA	Age 15 in grades 10-12	40 (10 <sup>th</sup> grade is modal grade for 2012) 3 (11 <sup>th</sup> grade is modal grade)	Math ( Reading Science	3 years	Group administration by xx in school; optional computer-based administration	Scores Performance levels
	TIMSS Advanced	Grade 11		Math Physics	4 years	Group administration	Scores Performance

## Annex A: Features of Existing International Large-Scale Assessments

Educational level	Assessment	Age/grade	Countries	Content domains	Periodicity	Administration	Results
						in by xx in school; optional computer-based administration	Score levels
TVET	None						
Higher Education	None						
Youth and adults	PIAAC	Adults age 16-65	33 countries	Literacy Numeracy Reading Problem solving in technology-rich environments	Not fixed	Individually administered at home	Scores Performance levels
	STEP	Adults age 15-64	12 countries	Literacy Numeracy	Not fixed	Individually administered at home	Scores
	LAMP	Adults age 15+	5 countries	Literacy Numeracy	Not fixed	Individually administered at home	Scores
	PISA for Development	15-year-olds		Math Reading Science	Once (2015-2016)	In school and out-of-school	Scores Performance levels
All learners	None						

## Annex B: Mathematics performance standards, TIMSS and ERCE

TIMSS PERFORMANCE levels for math (grade 4)	ERCE PERFORMANCE levels for math (grade 3)
<p><b>Advanced International Benchmark:</b> Students can apply their understanding and knowledge in a variety of relatively complex situations and explain their reasoning. They can solve a variety of multi-step word problems involving whole numbers, including proportions. Students at this level show an increasing understanding of fractions and decimals. Students can apply geometric knowledge of a range of two- and three-dimensional shapes in a variety of situations. They can draw a conclusion from data in a table and justify their conclusion.</p>	<p><b>Level IV.</b> Students can recognize the role governing the formation of a numerical sequence and identify its formulation. They can solve multiplication problems involving one unknown or that require making use of equivalency between the usual measures of length. They can identify an element on a two-dimensional plane and the properties of the sides of a square or rectangle to solve a problem.</p>
<p><b>High International Benchmark</b> Students can apply their knowledge and understanding to solve problems. Students can solve word problems involving operations with whole numbers. They can use division in a variety of problem situations. They can use their understanding of place value to solve problems. Students can extend patterns to find a later specified term. Students demonstrate understanding of line symmetry and geometric properties. Students can interpret and use data in tables and graphs to solve problems. They can use information in pictographs and tally charts to complete bar graphs.</p>	<p><b>Level III.</b> Students can solve multiplication problems or addition problems that involve an equation or require two operations. They can solve addition problems using units of measure and their equivalents or problems that include common fractions. They can recognize the rule governing a graphic sequence or additive numerical sequence and continue it. They can identify elements of unusual geometric figures and interpret the different types of figures for extracting information and solving problems using the data.</p>
<p><b>Intermediate International Benchmark</b> Students can apply basic mathematical knowledge in straightforward situations. Students at this level demonstrate an understanding of whole numbers and some understanding of fractions. Students can visualize three-dimensional shapes from two-dimensional representations. They can interpret bar graphs, pictographs, and tables to solve simple problems.</p>	<p><b>Level II.</b> Students can recognize the decimal and positional organization of the numbering system and the elements of geometric figures. They can identify a path on a plane and the most appropriate unit of measure for measuring an attribute of a known object. They can interpret tables and charts in order to extract and compare data. They can solve addition or multiplication problems involving proportions in the field of natural numbers.</p>
<p><b>Low International Benchmark</b> Students have some basic mathematical knowledge. Students can add and subtract whole numbers. They have some recognition of parallel and perpendicular lines, familiar geometric shapes, and coordinate maps. They can read and complete simple bar graphs and tables.</p>	<p><b>Level I.</b> Students can recognize the relationship of order between natural numbers and common two-dimensional geometric figures in simple drawings. They can locate relative positions of an object in a spatial representation. They can interpret tables and graphs in order to extract direct information.</p>