Learning assessments

BRIEF 2

In order to improve students' learning, we need to know more about how much students are learning now and where improvement is most needed. The assessment of learning is therefore a crucial issue that countries will need to strategically address as they develop education sector monitoring plans.

Learning assessments for diverse contexts and purposes

There are multiple kinds of assessments of students' learning, used in different contexts and for different purposes. Often we differentiate between two major categories: 1) individual assessments for pupils, and 2) system level assessments or evaluations for schools, regions, or national education systems.

Individual assessments

Individual assessments for pupils can be formative and give feedback to pupils and teachers on their skills and progression, or they can be summative, in the form of final grades or examination results.

In classrooms, teachers may design formative or summative tests to evaluate whether students are following the curriculum. Formative tests are diagnostic in nature: teachers want to know if learning is taking place and, if it is not, to provide appropriate interventions. Formative tests are also important for feedback to pupils and parents about the pupils' progression. Tests can also be summative, conducted at the end of the unit, term, or year, to determine whether students have acquired the required knowledge and skills. Teacher-designed tests are generally used as an assessment tool within a classroom or grade. They do not compare student learning across schools.

Public examinations have different aims to class-based tests. The results are typically used to certify that individual students have attained a certain level in their studies. An examination can also be used to assess whether or not schools are implementing the curriculum and whether teachers are delivering appropriate instructions. They may also be used to select students for further education. When the student's educational or professional future is dependent on their performance in an examination, it is referred to as a 'high-stakes' assessment.

System level assessment and evaluation

Large-scale regional, national, and international assessments are used to evaluate the output of a school system. These are instruments designed to provide evidence about the levels of student achievement in specific learning domains. Unlike examinations, which focus on individuals' results and determine certification or selection, assessment results have no consequences for individual students. Rather, the purpose is to assess how well students within the system are learning, and explain why some students are performing better than others. A variety of background information about students' teachers and learning environments are collected, along with data on students' learning outcomes. These are then analysed and linked together so as to provide informed policy suggestions that can be used by policy-makers.

Incorporating assessment data into a monitoring framework

Assessment data can play a key role in a country's overall monitoring framework, as part of the analysis of issues in an education system, and in order to monitor plans for improvement. In general, assessment data is most useful for monitoring purposes when the assessment has been rigorously designed for comparing different students, achievement levels, and time periods. National and regional examinations data may also be incorporated into a monitoring system through inclusion in the Education Management Information System (EMIS) or by making this information public via school report cards—though some caution must be exercised in using this information to analyse trends over time, since such examinations are usually not designed with that purpose in mind.

Available international assessment tools

Various international initiatives have created tools for education assessment in learning and education/school management. These international assessment tools are sample based, and are designed to provide evidence-based feedback to inform better education policy-making, leading, in turn, to improved learning and teaching. Broadly we differentiate between internationally developed tests to be used for system level analyses within countries, and large-scale international tests used for comparative assessment across countries. The large-scale comparative assessments are designed for comparing results on a common scale across countries, and they can also measure trends in learning outcomes over time.

It is important to note that the tools described below are not mutually exclusive. Some tools are multifaceted, generating data that can be analysed for various purposes. For example, SACMEQ data can be used for education planning, as well as for monitoring.

Assessments for Use at the Country Level

For use at country level, there are two internationally recognized assessment tools for testing of early reading and mathematics skills. These are the Early Grade Reading Assessment (EGRA) and Early Grade Mathematics Assessment (EGMA), developed by RTI International and funded by USAID. In addition, there are citizen-led initiatives such as UWEZO that carry out learning assessments at the household level.

EGRA – Early Grade Reading Assessment

This is an oral assessment tool designed to measure basic skills for literacy in children in early

grades. It focuses on the individual child and is a one-to-one assessment. The tool measures recognition of letters, reading simple words, understanding sentences and paragraphs, and comprehension. The assessment is adopted for use by a particular country in a given language. EGRA helps to establish national reading performance and the level of children's reading skills at an early stage, data which then feed into measures for improvement and policy-making.

EGMA – Early Grade Mathematics Assessment

This tool is the maths and numeracy equivalent of EGRA. It measures children's skills in numeracy and mathematics. It focuses on the foundations of maths, such as number identification, quantity discrimination (larger and smaller), missing-number identification, word-problem solving, addition and subtraction, shape recognition, and pattern extension. The assessment is crucial for determining ability for further numeracy tasks. It helps teachers to establish students' level of understanding of foundational skills and to identify areas of improvement towards further tasks in upper grade.

UWEZO – A civil society assessment tool used in Kenya, Tanzania, and Uganda

Uwezo, which means 'capability' in Swahili, is an initiative that carries out an annual household survey to assess whether children aged between 6 and 16 years have the standard literacy (reading) and numeracy (maths) skills required at Level 2. The assessment is regional, covering Kenya, Tanzania, and Uganda. The survey's assessment tools capture other demographic data, such as household income, location, and schools. They thus provide results that are broad and can be used in various education policy areas, giving room for robust analysis of, for example, schooling status, enrolment, attendance and teacher-student ratios. Also see ASER (India and Pakistan), Beekungo (Mali), Jangandoo (Senegal), MIA (Mexico), and PAL (the People's Action for Learning Network).

International large-scale assessments

Large-scale international assessments are designed to give relevant policy information on learning outcomes in a way that is comparable across educational systems. Typically these tests are sample based and consist of written or computer based cognitive tests, accompanied by surveys to pupils and principals. Some studies also include surveys to teachers and parents as well. Large-scale international assessments give countries and education systems the opportunity to benchmark against each other because results are presented on a common scale across countries. In addition they are designed to give reliable trend data for learning outcomes over time. This enables participating countries to assess strengths and weaknesses in their education system, and to judge the impact over time of reforms and education policy decisions.

A number of different international large-scale assessment tools exist, developed by different international organisations:

The Organisation for Economic Co-operation and Development (OECD) is responsible for PISA and PISA for Development, which are tests of learning outcomes. In addition the OECD also organises TALIS, which is an international survey directed to teachers and principals. The OECD has a broad mission to promote policies that will improve economic and social well-being, including through generating comparative data about different countries' education systems.

PISA – Programme for International Student Assessment

The Programme for International Student Assessment (PISA) is a triennial international survey run by

the OECD. PISA aims to evaluate education systems worldwide by testing the skills and knowledge of 15-year-old students. Every three years students from randomly selected schools worldwide take tests in the key subjects: reading, mathematics and science, with a focus on one subject in each year of assessment. To date, students representing more than 70 economies have participated in the assessment. PISA tests are not directly linked to the school curriculum. The tests are designed to assess to what extent students at the end of compulsory education, can apply their knowledge to real-life situations and be equipped for full participation in society. The information collected through background questionnaires also provides context which can help analysts interpret the results.

PISA for Development

This initiative aims to increase developing countries' use of PISA assessments for monitoring progress towards nationally set targets for education improvement, with a focus on student learning outcomes. It is also designed for tracking international education targets in the post-2015 framework. A pilot study is currently underway. The results will contribute to the post-2015 education-related development agenda.

TALIS

TALIS was established in 2008 as an international, large-scale survey of the teaching workforce, the conditions of teaching, and the learning environments of schools in participating countries. The study aims to provide timely, comparable, and useful policy information regarding the conditions of teaching and learning environments to participating countries.

The International Association for the Evaluation of Educational Achievement (IEA) is an independent, international association of national research institutions and governmental research agencies. IEA conducts large-scale comparative studies of educational achievement and other aspects of education, with the aim of gaining in-depth understanding of the effects of policies and practices within and across systems of education.

TIMSS

TIMSS (Trends in International Mathematics and Science Study) measures trends in mathematics and science achievement at the fourth and eighth grades. TIMSS has been conducted every four years since 1995. TIMSS reports overall achievement as well as results according to four international benchmarks (advanced, high, medium, and low), by major content domains (e.g., number, algebra, and geometry in mathematics, and earth science, biology, and chemistry in science. In addition the study collects information about curriculum and curriculum implementation, instructional practices, and school resources.

For countries where students are still developing fundamental mathematics skills, IEA's new TIMSS Numeracy assessment (designed to be administered at the fourth, fifth, or sixth grade) concentrates on measuring children's numeracy learning outcomes, including fundamental mathematical knowledge, procedures, and problem-solving strategies. There is also an advanced tool (TIMSS Advanced) that measures trends in advanced mathematics and physics for final-year secondary-schools students.

PIRLS and PIRLS Literacy

PIRLS (Progress in International Reading Literacy Study) is an assessment of pupils' reading comprehension and provides internationally comparative data about how well children read at the

end of grade four. PIRLS has been conducted at five-year intervals in countries around the world since 2001. In addition the study also collects information about home support, instructional practices and school resources in each participating country. Initiated in 2011, PIRLS Literacy (earlier known as prePIRLS) is based on the same view of reading comprehension as PIRLS, but is designed to test basic reading skills for countries where most children are still developing fundamental reading skills. PIRLS Literacy can be administered at the fourth, fifth, or sixth grade, and gives countries the opportunity to benchmark against the regular PIRLS test.

ICCS

The International Civic and Citizenship Education Study (ICCS) investigates the ways in which young people are prepared to undertake their roles as citizens. The study assesses pupils at the end of grade eight, and was last conducted in 2009. The study is again planned in 2016. ICCS reports on students' knowledge and understanding of concepts and issues related to civics and citizenship, as well as their beliefs, attitudes, and behaviours.

ICILS

The International Computer and Information Literacy Study (ICILS) is an international comparative study which is designed to evaluate students' ability to use computers to investigate, create, and communicate in order to participate effectively at home, at school, in the workplace, and in the community. The study is aimed at students at the end of grade eight, and was first established as a baseline study in 2013 with 21 participating education systems around the world. The next round of ICILS is planned in 2018.

There are a number of regional assessment programmes, such as SAQMEC and PASEC in Sub-Saharan Africa, and LLECE in Latin America, and other newer initiatives. These programmes aim to monitor and evaluate school systems and to provide evidence-based information that can be used by policy-makers to plan and improve the quality of basic education.

SACMEQ – The Southern and Eastern Africa Consortium for Monitoring Education Quality

This is an umbrella organization for 16 education ministries in Southern and Eastern Africa. The organization brings these ministries together to share experiences and expertise with a view to the scientific monitoring and evaluation of education policies on school conditions and the quality of education. SACMEQ carries out training programmes to equip education planners with technical skills, including data collection and analysis for monitoring and evaluation purposes. In addition to its focus on monitoring and evaluation, SACMEQ also occasionally carries out reading and math assessments in member countries to assess sixth-grade students' abilities in mathematics and reading English.

PASEC – Programme for the Analysis of Education Systems

PASEC is a regional assessment tool for Francophone countries in West Africa and Asia, conducted by CONFEMEN (La Conférence des ministres de l'Éducation des pays ayant le français en partage). It provides information about the performance of education systems, which contributes to the development and monitoring of education achievement in member countries. In addition, PASEC carries out comparative evaluations across its member countries. The objectives of PASEC are to: assess pupils' performance and identify efficiency and equity issues for basic education; provide national policies with indicators that will allow them to make relevant comparisons; foster, at national level, the development of an internal and permanent capacity for the evaluation of the education

system; disseminate evaluation results at international level to contribute to reflections and discussions on factors determining education quality. The overall aim is to contribute to effective evaluation methods that will enhance the capacity of national ministries to evaluate learning achievement at primary-school level.

LLECE – Latin American Laboratory for Assessment of the Quality of Education

LLECE is a network of national units focusing on the assessment of education quality. It was initially created with 15 members—Argentina, Bolivia, Brazil, Chile, Colombia, Costa Rica, El Salvador, Guatemala, Honduras, Mexico, Paraguay, Peru, Dominican Republic, Uruguay and Venezuela, coordinated by the Regional Bureau of Education of UNESCO for Latin America and the Caribbean. The network arose in with objectives of providing more information for designing appropriate educational reforms, sharing and developing the expertise necessary for conducting educational assessments, and having a more open orientation toward the sharing of assessment data with the public. So far, three regional assessments have been conducted: PERCE (1997-98), SERCE (2006-8), and TERCE (2013-15). From the initial focus on only reading and mathematics, the approach has grown to also include writing and the natural sciences (the latter only for pupils in the sixth year of primary school). It also aims to identify the factors associated with different levels of achievement, such as the socio-economic context, family life and personal issues, educational policies, and school processes.

As the importance of rigorous learning assessments gains increased recognition, other new regional and international assessment programmes may emerge. As one example, the Southeast Asia Primary Learning Metric (SEA-PLM) was recently inaugurated, in association with the Learning Metrics Task Force (LMTF), a global initiative that is working to improve the measurement of learning around the world.

Key considerations in designing national learning assessments

Reliable information on pupil performance is the key to the successful implementation of targeted education policies. In the past two decades, national assessments have emerged as an important tool for providing a measure of educational achievement. There exist a great variety of national assessment programs with different aims and purposes. Broadly, one can differentiate between assessments that are designed for accountability at all levels, and assessments that are designed for system evaluation and development. In both cases, the desire to measure change in achievement over time will imply specific requirements to the test design.

When designing a national assessment system, there are some key initial considerations that will guide all further choices for the development of the system:

- What is the intended purpose of the assessment?
- Which competencies do we want to test?
- Which are the main target groups to be tested?
- How can validity and reliability be ensured?
- Which format should the tests have?
- Is it important to measure trends, and if so how can this be done accurately?
- How should the results be reported and to whom?
- Do we have the necessary expertise?

These aspects are typically defined within a national assessment framework. Following are some

further considerations regarding each issue:

Purpose of the national assessment

A national assessment program can serve multiple purposes, and the main purpose should determine the design of the assessment. It is therefore very important to be clear from the beginning about the main purpose. The use of one single test for several purposes might be inappropriate as the information ideally required in each case is not the same. Therefore, education authorities are advised to rank the different purposes in order of priority and adjust test designs accordingly. (See Standards, Accountability, and Student Assessment Systems, Canadian Education Association (CEA))

There are three general purposes for most national assessments. The first group consists of tests which summarize the achievement of individual pupils at the end of a school year or at the end of a particular educational stage, and which have a significant impact on their educational careers. These are high stakes tests, which are often referred to as summative. Second are assessments intended to monitor and evaluate schools and/or the education system as a whole. In this case, test results are used as indicators of the quality of teaching and the performance of teachers, but also to gauge the overall effectiveness of education policies and practices. A third category is composed of assessments that are mainly for the purpose of assisting the learning process of individual pupils by identifying their specific learning needs and adapting teaching accordingly.

Competencies to be tested

The assessment domains can either be based on particular subjects in the curriculum, or can test key competencies for learning across subjects, such as numeracy, literacy, problem solving, or information and communication skills. The assessment of key competencies will be most relevant for formative assessment programs designed to monitor education systems and/or identify individual learning needs. All national assessments measure cognitive skills in the areas of language/literacy and mathematics/numeracy, a reflection of the importance of those outcomes for basic education. In some countries, knowledge of other areas, such as science, social studies, particular languages or other domains, are also included in an assessment.

Whatever the domain of the assessment, it is important to develop an appropriate framework that clearly defines the competencies and skills to be tested, and a test specification. This is necessary both for constructing assessment instruments and afterward for interpreting results.

Target groups and Sampling strategy

The selection of target groups for the assessments is dependent on the purpose of the test. If the purpose is mainly formative, the tests should be administered at stages where the acquired competencies are crucial for further learning and development. Examples can be the beginning and end of primary and towards the end of lower secondary education. If the main purpose of the assessment is summative, it would typically be performed at the end of an education level, such as upon the completion of primary school, lower secondary, or upper secondary.

Examinations and tests to monitor schools are often compulsory for all pupils, while tests that concentrate on evaluation of the educational system as a whole are often administered to a representative sample. If a test is sample based, it is necessary to consider how the results are to be

reported when the sample is defined. If results are to be broken down by regions, school types, gender, language of instruction etc., one has to make sure that the sample is representative at all those levels.

Validity and Reliability

Test validity is the extent to which a test actually measures what it is intended to measure. Validity is generally considered the most important issue in educational testing because it concerns the meaning placed on test results, and the extent to which the results of the test can be trusted to have measured the right competencies. A highly valid assessment ensures that all relevant aspects of student performance are covered by the assessment. There are statistical methods to calculate a test's validity.

Test reliability is the degree to which an assessment produces stable and consistent results. Adequate reliability is a necessary condition for the validity of a test. This means that if the measurement is not reliable it cannot be valid either. Newer scaling methods (Item Response Theory or IRT) have resulted in a different understanding of test reliability, due to the recognition that individual items may differ in their difficulty level. When using IRT methods, test reliability roughly means the precision of the measurement at different levels of the competency measured. The converse of reliability is measurement error and therefore it is of utmost importance to ensure the best possible reliability of the test as a whole.

Test design

To ensure the validity of the test, it must consist of test items representing the whole range of the test domain described in the framework. The test must also contain enough items for each proficiency level. Items can be either multiple-choice or open ended, or a combination of both. Open ended questions require a very strict scoring manual and schooling of scorers, however. In many countries there is now a rapid movement from paper based towards computer based testing. This opens up the possibility of adaptive testing, where the test is automatically adjusted to the student's proficiency level. This method allows for more precise measurement of the whole competency and more targeted testing.

A rotated test design (matrix sampling) is often used for sample based tests to monitor a whole education system. In a rotated design, the test is constituted on a set of booklets or in blocks each representing only a part of the whole test. Each student only answers one booklet, which can contain different blocks of material. This enables testing of a large set of items without making the test too long for each student. However, with this method it is not possible to deliver individual results for students.

All test items for any of these types of tests must be piloted and analysed using psychometric methods before they are used in the final test, in order to make sure that the test meets all requirements for validity and reliability.

Measuring trends

To measure development of learning achievement over time, the test must contain a set of anchoring items, which are repeated every cycle. The anchoring items will be used to make sure the reported proficiency levels represent the same level of difficulty over time. In other words that the numerical

results always represent the same level of competency. Anchoring items must be kept confidential to ensure the same test conditions over time. Only by using a test design of this type can trends be monitored.

Reporting results and using scaled scores

Ideally, assessment results should be reported both to decision-makers and to the general public. The public sharing of such information helps ensure that all stakeholders can help hold educational institutions accountable. However, the public sharing of data that has been disaggregated down to the school level can raise many controversies and is not always beneficial; in some cases higher levels of regional aggregation may be more appropriate. Additionally, if school-level reports are made public, extra care must be taken to ensure that individual students cannot be identified.

It is often necessary to prepare more than one type of report—some more detailed than others—in order to ensure that key findings are accessible to different audiences. Depending on the test design and the purpose of the test, the results can be reported either as one total test score, or be broken down in sub-scales representing different sub-domains and proficiency levels.

To achieve comparability, standardized testing programs most often report the outcomes as scaled scores. The reported scaled scores are obtained by scaling the raw scores (percentages or score points) onto a common scale to account for differences in difficulty across different forms.

Test expertise

Development of national tests requires high expertise, both curricular and content specific expertise and high psychometric competence. An important consideration is how to ensure relevant expertise during the whole process. In some countries there are national institutes or test centres that contribute the necessary expertise, but often this is not the case. There are, however, a number of national and international test institutes that will be able to support countries and provide important capacity building.