Designing a national assessment

BRIEF 3

Educational measurement

In the past two decades, national assessments have emerged as an important tool for providing a measure of educational achievement. Although there is a great variety of national assessment programmes with different aims and purposes, most seek to measure change in learning outcomes over time. National assessments can also be used to provide data to monitor national progress towards Sustainable Development Goal (SDG) 4. A national assessment programme requires detailed planning based on the key considerations listed below.

What is the intended purpose of the assessment?

According to the UNESCO Institute of Statistics (UIS), most national assessments have two general purposes:

1. ‘to monitor to what extent students are reaching key learning objectives as outlined in national curricula and to support learning for all’ and
2. ‘to hold schools accountable and to provide students and their parents with information about learning progress’. (UIS, 2017: 7)

Which competencies will be tested?

Assessments can be designed to test general competencies across subjects (such as literacy, numeracy, problem-solving, or communication skills), as in the Programme for International Student Assessment (PISA), or to measure the intended or achieved curriculum, as in other instruments. All national assessments measure cognitive skills in language/literacy and mathematics/numeracy, with some countries’ assessments also covering other domains such as science, social studies, and languages. Assessing these varied competencies may require the use of different assessment instruments, such as oral, practical, or portfolio components.

Whatever the domain of the assessment, it is important to develop a framework that clearly defines the competencies and skills to be tested. One of the challenges associated with assessing cross-curricula competencies is that it can be difficult to reach agreed definitions of these skills. In such cases, external experts may need to be brought in to support the task of competency definition.

To understand the variables that may affect learning there needs to be careful consideration of what background information should be collected and how it should be gathered (e.g. via teacher or
Who are the target groups of the assessment?

When selecting a target group, countries should consider if the assessment should:

- Target an **age group or a grade level**? There are advantages and disadvantages to both methods. One benefit of grade-based sampling is that it allows for more background information on teaching practice and classroom conditions to be linked to learning outcome data.
- Be **sample-based or census-based**? While examinations and tests to monitor schools are often compulsory for all pupils, tests that concentrate on evaluation of the educational system as a whole are often administered to a representative sample.
- Include **out-of-school children**? For example, by using household surveys.

How can the quality of the assessment instrument be assured?

A quality assessment is characterised by content, concurrent and predictive validity, reliability, and fairness.

**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 be measures of the right competencies. A highly valid assessment is one that covers all relevant aspects of student performance. Methods to estimate a test’s validity include cross-validation, item analysis, inter-correlation of items, and factor analysis.

**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: if the measurement is not reliable, it cannot be valid. Newer scaling methods, Item Response Theory (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, therefore accuracy and precision of measurement is of utmost importance in ensuring the best possible reliability of the test as a whole.

**Fairness** of an assessment refers to its freedom from any kind of bias. Any test should be appropriate for all respondents, irrespective of race, religion, gender, or age. An assessment should not disadvantage a respondent on any basis other than on their lack of knowledge and skills the assessment is intended to measure. To ensure that a test meets requirements for validity, reliability and fairness, test items must be piloted and analysed using psychometric methods before they are used.

Which format should tests have?

To ensure validity, a test must consist of test items representing the whole range of the test domain. 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 thorough training of scorers. Many countries are moving from paper-based towards
computer-based testing. This opens up the possibility of adaptive testing, where a test is automatically adjusted to the student’s proficiency level thus enabling more precise measurement of the whole competency, and 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 in blocks, often in a set of booklets, each block representing only a part of the whole test. Each student answers just one booklet, which can contain any one of the different blocks of material. This enables testing of a large set of items without making the test too long for each student. However, this method does not allow individual students’ results to be delivered.

How can performance trends over time be measured accurately?

To monitor trends of learning achievement over time, the test must contain a set of anchoring items, which are repeated every cycle. Anchoring items can be used to ensure the reported proficiency levels represent the same level of difficulty over time—in other words, that the numerical results always represent the same levels of competency. Anchoring items must be kept confidential to ensure the same test conditions over time.

Who should implement the test, and how frequently?

It is important to consider how an assessment will be implemented. Countries may consider whether the assessment should be implemented by a government ministry or an independent specialist group, and whether it should be administered by a trained external administrator or teachers. The purpose of the assessment should also determine how frequently it is administered and its timing in the school year, for example at the beginning or the end of the school year.

How should the results be reported and to whom?

Assessments should answer key policy questions, and results should be reported to both decision-makers and the public. However, public sharing of data that have been disaggregated down to the school level can be controversial. If school-level reports are made public, care must be taken to ensure that individual students cannot be identified. It may be necessary to prepare several reports – some more detailed than others – to present the findings to different audiences such as policymakers, teachers, or the public. The results of international assessments are usually published as national reports. Depending on the test design and purpose, results can be reported either as one total test score, or broken down in subscales representing different subdomains and proficiency levels.

Is the necessary expertise available?

Development of national tests requires both curricular and content-specific expertise, as well as psychometric competence. Some countries have national institutes or test centres that contribute the necessary expertise. There are also national and international test institutes that can provide country support and capacity building.

Tools
World Bank. *SABER: Student assessment*.

References and sources


