CHAPTER 4

Using Student Learning Outcomes to Measure Improvement

As discussed previously, student learning and growth over time are key criteria against which educational systems, local education authorities, schools and teachers are to be held accountable. An important challenge, therefore, is to properly assess student learning and growth. A single type of assessment cannot fully reflect student learning. All forms of assessments, from standardised tests to portfolios of students' work have issue of validity, reliability and objectivity (Baker, 2010). It is important to develop a system that uses different measures of student achievement and multiple sources, in which assessment data can serve as a quantitative anchor (OECD, 2010).

The first section of this chapter provides an overview of the options for assessing student learning, highlighting the strengths and weaknesses of the different methods, with relevant international examples. Based on the discussion of the formative and summative assessment options, the section suggests that educational systems such as Australia, Alberta (Canada) and Hong Kong-China benefit from having different sources of information regarding student performance to ensure the highest level of completeness and accuracy.[1] Having a battery of valid, reliable and varied measures of student learning and growth, however, also has clear implications in terms of costs and capacities required, particularly at local levels. Not surprisingly, the majority of OECD and partner countries apply student assessments based on tests of student achievement (OECD, 2008).[2] Assessments of student learning are used in different countries for a range of purposes, including for gauging the performance of the system as a whole, for diagnostic purposes tied to improvement efforts (*e.g.* Mexico), for accountability (*e.g.* the United States and United Kingdom), for incentives for teachers and schools (*e.g.* Chile), and for combinations thereof (OECD, 2009a). As an integral part of an educational system, assessments themselves are reviewed, evaluated and modified in OECD and partner countries to better reflect policy priorities, education reforms in related areas such as curriculum, and the demands of a rapidly changing world (*e.g.* Australia, Brazil, Norway, the United Kingdom and the United States).

Although there is no single model of assessment that can be gleaned from international practice, some of the technical, logistical and political challenges are common across education systems. Some of these common issues and some of the recommended practices in terms of linking assessments to standards and curriculum, are presented in this chapter. Based on a country's policy priorities and the conditions, constraints and opportunities of a particular education system, the challenge will be to find the right combination of different assessments, their relative weighting, and their uses and consequences (OECD, 2009b).

Developing these different complementary methods of student assessments takes time and resources. Education authorities can establish a gradual process that takes advantage of the immediately available sources for school improvement and accountability initiatives, while also having a longer-term vision of assessment. Because of the cost-effectiveness of standardised student assessments, as well as the relative comparability of results across diverse national contexts, externally administered standardised assessments are used in several OECD and non-OECD countries for both accountability and improvement purposes. The ENLACE assessment in Mexico, begun in 2006, and the *Prova Brasil* that the Brazilian federal government implemented for the first time in 2005 (Box 4.1), offer good examples of dynamic development (Zúniga Molina and Gaviria, 2010; Parandekar *et al.*, 2008). Other examples across OECD countries show that it is possible for education systems to implement external assessments, while allowing education practitioners to innovate in their practice (OECD, 2009). As a specific example of the opportunities for standardised assessment to allow innovative practices, the chapter provides an overview of the main characteristics of the ENLACE assessment in Mexico, and concludes with specific considerations and recommendations for educational authorities for its further development.

4.1 STUDENT LEARNING OUTCOMES: ASSESSMENT INSTRUMENTS AND MEASURES

Student results, whether actual scores or marks, or the proportion of students attaining specific and pre-determined performance levels, may be based on one or more measures of student learning. These may involve student essays, extended projects, portfolios of student work, multiple choice or short answer tests, among others, and may be used for formative or summative assessment (OECD, 2009; Baker, 2010).

Box 4.1 *Prova Brasil* for accountability and improvement

Brazil's first census-level student assessment ***Prova Brasil*** was administered by the Ministry of Education for the first time to test proficiency in mathematics and Portuguese in 2005. The assessment is administered every three years to primary and secondary students in Grades 4 and 8, and represents one of the government's main efforts to establish a results-oriented accountability framework focused on student achievement. A recent study conducted by the Ministry of Education used student achievement data from the assessment and regression analysis to identify municipalities with superior performance, even after considering students' family and socio-economic background. Using qualitative methods, such as classroom observation and interviews, the study further attempted to identify good policies and good practices at the local level that may be contributing to superior performance.

Further information is available (in Portuguese) at *http://provabrasil.inep.gov.br/*.

Source: Moriconi, 2009; Parandekar *et al.*, 2008.

It is common to suggest that assessments should best represent the cognitive demands or thinking skills that are considered most important by the education system. Constructed student responses are often considered preferable because students have to reach into their repertoires, search and then apply their learning (Baker, 2010). Likewise, for measures addressing skills either easily memorised or easily developed outside of school (*e.g.* via the Internet), education systems may choose more efficient and cost-effective testing processes, saving extended and more expensive assessments for learning that requires difficult understanding, applications and communication of rich or complex content (Baker, 2010).

Technically, however, there is not a bi-univocal relation between the cognitive level of the skill to be assessed and the type of items to be used (Zúniga Molina and Gaviria, 2010). An important consideration is the amount of information provided by a specific item. A single constructed response item can contain a higher amount of information, if carefully stated, than the corresponding multiple choice question. A series of multiple choice items, however, designed to the same cognitive demand, could provide, if carefully designed, the same amount of information. A trade-off exists, therefore, between the information provided by each particular item and the facility to automatically mark the responses given by the test takers.[3] Often decisions regarding which instruments and measures to employ are made on the basis of cost and feasibility rather than on optimal assessment for particular standards (Baker, 2010). To compare approaches, Table 4.1 presents a summary of some of the main characteristics, strengths, limitations, cost implications and technical issues of different assessment options.

There are a limited number of high-quality approaches to ensure the comparability of performance assessments, particularly if the interest is in providing some degree of feedback regarding the teaching and learning process. If not properly designed, assessments may not shed sufficient light on which aspects are well or poorly learned, thus providing little or no guidance for system improvement. Performance assessments may be developed using relatively clear domain boundaries for content and cognitive demand, in which case comparability may be easier to establish. Because any set of constraints limits the range of student performance that can be assessed, including tasks that do not have a specific focus could determine the degree to which students can transfer their learning to new situations (Baker, 2010). The design of the PISA assessments, for example, follows this logic. The degree of transfer may be relatively small (but nonetheless difficult), for example where students are asked to perform a mathematics procedure presented in a previously unseen format. Transfer may be more difficult when students are given a problem requiring the application of different strategies rather than a common procedure for solution (Baker, 2010).

Table 4.1

Instruments and sources of evidence to assess student learning

Assessment format	Comments/Uses	Strengths	Weaknesses (sources of bias, validity and reliability issues, cost-considerations, and capacity requirements)
Long open-ended responses, projects, essays	Formative, summative or combined use. Curriculum embedded; may require teacher assistance.	Task validity, rich content, high cognitive demands; transfer and application a possibility.	For accountability: cost and comparability; replacement costs; training of teachers for valid and reliable scores; weak alignment with standards.
Portfolios of student work	Classroom use; weak evidence for accountability, except selection to higher programmes; may include required or chosen elements in the same or different content.	Assessment over time, showing progress or flexibility on multiple topics; choice of topics, style or content; transfer and application possible.	Scoring unreliability if student choice is offered; comparability among students; cost of scoring if external to the classroom; conflict for teachers if used in accountability; requires extensive teacher training; if used for accountability, high cost.
Classroom observation	For use in teacher effectiveness or as opportunity to learn explanation for outcomes. May be conducted by peers in school, other teachers, administrators, or pedagogical or content experts.	Real time sense of teacher and student activity; feedback for teacher; may be conducted by peers; value to explain data on student or value-added modelling reporting; feedback for teacher evaluation.	Requires agreement on learning model and relationship to standards; need multiple visits; trained observers; high and low inference rating scheme; if purpose is feedback, training to observe and give feedback validly and reliably; observation biases teacher and student behaviour; not easily scalable unless random samples of video with significant scoring costs.
School- and class-based tests of any format	Strengthens instruction and alignment; with quality control builds repertoire of assessment events and instructional interventions; may be combined with summative assessment.	Fits the curriculum as taught; immediate feedback to students and intervention possible; builds teacher capacity; provides new examples for outcome examinations.	May be closely or loosely related to standards; may be of poor quality (psychometric characteristics); scoring schemes may not be explicit. Training in assessment design, administration and scoring required. If used for formative assessment, strategies for improvement required; if used as part of the accountability system teacher conflict of interest is possible.
Standardised assessments using multiple choice and short answer *(externally provided)*	Used commonly for broad summative purposes (diagnostic and accountability-focused).	Inexpensive to score; good for vertical equating and growth modelling; reliable. External character decreases sources of bias stemming from local relations.	Validity is a question if not properly tested; carefully designed content and cognitive demands may be shallow; alignment to content and curriculum may be weak. May encourage teaching to the test and other non-desirable behaviours; limited transfer of knowledge and skills to applied settings.

Source: Baker, 2010.

Marking

Open-ended or constructed student responses are commonly marked by teachers, the students' own or by teams of teachers specially trained to mark examinations (Baker, 2010). Training may occur by having markers examine a range of student responses against a set of pre-validated or expertly scored examples. In some cases, a chief examiner may prepare the paper, and marking is based on deviations from the model. In other cases, training involves exposing markers to the variety of ways students may achieve a score level. In all cases, data are captured about the effectiveness of the training; usually by requiring the teacher to mark a set of papers at a level that is considered adequate to qualify. Some marking training sessions emphasise reliability, focusing on the degree to which individual teachers agree with one another. This approach, when used without validation, may lead the markers to define quality in terms of socially shared expectations, and as one group of markers may differ from another, resulting in varying levels of stringency used for different groups of students. This can undermine the trustworthiness of the results. With adequate quality training, and with common scoring dimensions, the reliability of markers and the validity of their marks can be high (Baker, 2010).

Logistics, costs and technology

Training to mark an examination consisting of extended student work may take between four hours and two days (Baker, 2010). Many training sessions are followed by marking sessions, so costs include the cost of the markers (if teachers, their typical daily work time), travel, housing and food. In an effort to contain costs, common training may be conducted in a set of centrally located sites or remotely by computer, with telephone or chat support either mandated or available. The actual marking may be done at home, with no supervision, but with access to or mandated calls by the marking supervisor. Outsourcing examination marking to non-teachers or teachers in other locales has been used in some educational contexts, but it is not common (Baker, 2010). Individuals can be trained either in person or remotely. When marking is done without supervision (*e.g.* at the teacher or marker's home), the costs of marking can be significantly reduced, since no payments for travel or other on-site expenses are required, and compensation may be on a piecework basis, depending upon the number of papers marked. In some countries and sub-national jurisdictions, marking papers is a routine expectation and included in any collective bargaining conducted by the teachers' union (*e.g.* Alberta, Canada), while in others, marking of assessments "need not be done by teachers".[4]

More recently, technology has been used to score student essays with a reasonable degree of success. Approaches involve the use of pre-marked essays and a complex regression model that includes linguistic and lexical aspects of the students' work (Burstein, 2003). Some of these approaches require time-consuming training and the rating of papers in advance of the computer marking, a procedure that must be carried out for each and every change in topic. Other approaches work in a manner similar to grammar and spelling checkers in word processing software. Complex natural language understanding systems are also available, but to date these still require intensive work to adapt them to different topics and different levels of student work (Baker, 2010). Computer scoring of problem-solving and other open-ended responses is being developed (Chung and Baker, 2003; Chung *et al.*, 2001), and is most useful if the problem has a specific set of right answers (Baker, 2010). Computer scoring can also map (through neural networks) the paths taken by different students to achieve success. The latter data are useful for formative assessment (Baker, 2010).

Two of the primary limitations, however, for computer marking have been the availability of equipment and computer literacy of students and teachers. There have been significant improvements in optical scanning of student writing and voice recognition software. Within a few years, this may have important ramifications for the current paper-based assessment practice worldwide (Baker, 2010). The structural constraints of computer access, computer literacy and connectivity, however, remain challenges for emerging and developing economies (ITU, 2009).[5]

Quality of assessments

The technical quality of assessments is a major issue when findings are used to make high-stakes decisions for students, teachers, principals, or other individuals.[6] A key criterion regarding the technical quality of an examination is *validity*. Validity depends upon the purpose of the test and the evidence that the uses of the test are appropriate. If the purpose of the test is to assess accurately students' acquisition of content and skills, then inspection of the tasks or items and the estimate of depth of sampling are important considerations of content validity. Older notions of validity, including "face" or content validity, concurrent validity and predictive validity, have been subsumed in an overall consideration of validity for tests or examinations (Baker, 2010; Linn [ed.], 1993).

If the purpose of the test is to select candidates for higher levels of schooling, then the test used may well be examined with regard to its ability to predict success in further schooling. As accountability systems and the assessments within them have evolved, assessments appear to have multiple purposes, a fact that makes validity estimates more difficult. For example, if an accountability test is designed to place students accurately in a classification or on a continuum reflecting their level of mastery of subjects, then the assessment needs to be able to differentiate among students in different classes, asking more difficult questions of those expected to have more highly developed skills. If the test is intended to measure the effectiveness of the educational system, then validity evidence should be available to show that the test is sensitive to high-quality instruction. If the test performance does not change as a function of teaching, but rather by maturation or other non-school influences, it is not appropriate for use in accountability systems. In addition, results from the same examination may be expected to be used by teachers to revise their instructional sequences, either in the same school year or across years. There must be evidence, therefore, that the reported results provide sufficient and relevant information for that function. If tests are expected to monitor students' growth over a number of years, then the idea of vertical scaling (difficulty of tests is equivalent in different years) is essential. Such a requirement can also support value-added models that attempt to identify the contributions of schools in improving student outcomes (treated in more detail in Chapter 5).

Box 4.2 Mixed systems of student assessments

***Victoria, Australia:* Combining school-based and state assessments**

In the state of Victoria, Australia, the Victorian Curriculum and Assessment Authority (VCAA) has established the Victoria Essential Learning Standards to provide a yearly description of what all primary and secondary students are expected to learn and achieve. The VCAA also administers the National Assessment Program – Literacy and Numeracy (NAPLAN) and provides school-based, on-demand assessments for schools. Teachers are involved in developing school-based assessments, along with academic support staff, and all prior year assessments are available to the public. At least 50% of the total score for students consists of classroom-based tasks (*e.g.* lab experiments, investigation on key topics, and extended reports). At the same time, as part of the NAPLAN, approximately 260 000 students in years 3, 5, 7 and 9 undergo standardised assessments throughout Australia. The system thus combines school-based assessments with standardised state assessments.

Further information is available at *www.vcaa.vic.edu.au/.*

***Alberta, Canada:* Developing a holistic framework for assessment**

In Canada, the Alberta Education Authority commissioned the *Alberta Student Assessment Study*, a review of theory and practices relative to student assessments that provided recommendations on:

- Curricular learning outcomes, performance standards, and the reporting of student achievement.
- How external assessments and classroom-based assessments of student achievement can be used optimally to inform important decisions on student needs, school management and issues at the provincial levels relating to ensuring learning opportunities for all students.

Results from the study were presented in 2009, with specific guidelines and recommendations on how the education system could effectively combine performance standards, classroom-based assessment and provincial assessment, reporting of student achievement and professional development of teachers. Its recommendations are based on sound evidence and provide useful references for other systems looking for ways to establish complementary approaches to student assessment, within an accountability and improvement framework.

Further information is available at *http://education.alberta.ca/department/ipr.aspx.*

Sources: Victorian Curriculum and Assessment Authority, 2010; Darling-Hammond and McCloskey, 2008; Government of Alberta (Canada) – Education, 2009.

A review of test results should consider whether the examination is relevant to its purposes, and whether there is evidence to document the examination's ability to deliver on its purposes or claims. Getting validity evidence is difficult, especially during a developmental testing period when the prospect of public reporting or sanctions does not exist. Experiments can be conducted to determine whether assessments are sensitive to instruction, whether they properly categorise students who have done well on similar measures (concurrent validity), or whether relevant performance standards have been set at an appropriate level (as opposed to politically defined levels) (Baker, 2010).

The *reliability* of assessments is also a vital issue. This refers to the consistency with which a test measures performance. Reliability may naturally degrade if change (*i.e.* improvement) is desired. Using proper statistical controls may solve this problem partially. Measures used in accountability must also meet the challenge of *fairness*. Fairness does not mean equal outcomes, but that the characteristics of the examination and marking do not advantage any particular group, other than those most well prepared. For education systems that serve heterogeneous student groups (*e.g.* with varying socio-economic backgrounds, from different ethnic groups or with different languages spoken at home), fairness of assessments is an important issue. Linguistic features of students might confound estimates of learning in other subjects, like mathematics or the sciences. Other issues are that tasks may be relevant to only a subset of students (*e.g.* urban or rural) or show differences in performance independent of competence in other similar tasks in the domain. Weaving these technical requirements into a fabric that supports the technical quality of the measures used in accountability requires forethought and discussion with relevant stakeholders to determine which features may pose challenges for wide acceptance (Baker, 2010).

Because all measures of student learning, including assessments, may present potential shortcomings and sources of error and bias that can affect validity and reliability, education systems may opt to have different sources of information on student performance to ensure the highest level of completeness and accuracy (Baker, 2010). Assessments of student learning are used in different countries for different purposes. The challenge is therefore to find the appropriate balance of standardised assessments, school-based assessments, externally and internally marked and referenced, for different purposes and within the capacity, budget and structural constraints of the education system. The following section reviews an important student assessment in Mexico and discusses some of its main characteristics in light of the previous discussion in order to identify challenges and opportunities for its further development.

4.2 THE ENLACE ASSESSMENT SYSTEM IN MEXICO

In 2006, SEP implemented the first round of the annual National Assessment of the Academic Achievement in Schools (*Evaluación Nacional del Logro Académico en Centros Escolares*, ENLACE). ENLACE was designed to provide information to students, parents, teachers, principals and the general public regarding individual student achievement and grouped results at the school level.[7] In contrast with the EXCALE exams that are administered to samples of students in different grade levels,[8] the original purpose of ENLACE was to serve as a benchmark to inform improvements in teaching and learning processes at the school and classroom level for primary and secondary students (Zúniga Molina and Gaviria, 2010).

Mathematics and Spanish have been tested in every round since 2006, with a third subject varying each year: science was included in 2008, civics and ethics in 2009 and history in 2010; geography will be included in 2011. The exam is currently applied to primary students in years three to six and secondary students in years one to three. The test was applied to the first two years of secondary school for the first time in 2009. Overall, in 2009, the ENLACE assessment was taken by more than half (51%) of all students at the pre-primary, primary and secondary levels in Mexico (INEE, 2010, Indicator ED01; Zúniga Molina and Gaviria, 2010).

Levels of student achievement reflected in ENLACE results since 2006 are in general agreement with PISA 2006 results, that is they are low on average but there has been improvement. Between 2006 and 2009, for example, the percentage of students classified as "unsatisfactory" or "regular" dropped from 78.7% to 67.2%, while the percentage of students classified as "good" or "excellent" rose from 21.3% to 32.8%.

Results from the 2010 ENLACE application, as well as results from PISA 2009, will give further information on trends in student achievement.

Administered by SEP through the General Directorate for Policy Evaluation (*Dirección General de Evaluación de Políticas*, DGEP), the ENLACE assessment has become a socially accepted measure of student performance (Zúniga Molina and Gaviria, 2010; Salieri, Santibañez and Naranjo, 2010). The test is administered in April each year and results are presented publicly in September of the following academic year. DGEP processes test results using both commercial and proprietary software, and produces materials for users to interpret the results that are presented via school information packets and via the Internet *(www.dgep.sep.gob.mx)*. Through this website, students, families and any interested person can obtain information, using a special identification number on the student's answer sheet. The results can also be seen in aggregated form, by school,[9] by state or at the national level. Media coverage of the results is widespread and although the practice is discouraged by officials, different versions of "school tables" comparing grouped averages of raw scores of students by school is common practice. The importance that SEP and state education authorities have given the public presentation of results of the ENLACE assessment is supported by international comparisons. As discussed in Chapter 3, results from the PISA 2006 results indicate that the strongest impact on student performance across countries was related to the publication of schools' student achievement data (OECD, 2007).

The information that schools are supposed to receive through state educational authorities includes the proportion of students at each achievement level by grade and content subject. Each school should also receive information on the proportions of students at each achievement level compared with the results of the students and schools of the same type, at state and national levels. The information is organised so that it is useful for identifying possible teaching improvement opportunities and for allowing groups to compare their results against those of other schools with similar socio-economic conditions and infrastructure (Zúniga Molina and Gaviria, 2010).[10] Thus, teachers, school principals, and students and their families can assess progress and the difficulties encountered in learning, including identifying parts of the curriculum that have not been appropriately addressed. Teachers are expected to analyse students' results and identify strengths and weaknesses in the subject areas tested.

A recent review of state-level uses of the ENLACE results showed that they have become a national and local reference of students' learning achievement, with most state education authorities conducting some form of follow-up activities (Salieri, Santibañez and Naranjo, 2010). SEP provides all state educational authorities with printed brochures, reports and CDs to be distributed to schools regarding individual and school-grouped performance. Some states such as Jalisco, Nuevo León and Veracruz have developed their own materials that are distributed to supervisory staff (*Supervisores* or *Jefes de Sector*), and offer some form of support and professional development courses to schools identified as under-performing based on collective ENLACE results and needs assessments (Salieri, Santibañez and Naranjo, 2010). This underscores the importance of and opportunities for state educational authorities regarding improvement efforts and accountability mechanisms (treated in Chapter 7 and other chapters of the report).

Design and technical characteristics

Based on a consideration of the basic design elements, characteristics and test results since 2006, the ENLACE assessment instrument presents robust levels of internal consistency, validity and reliability as a measure of student learning (Zúniga Molina and Gaviria, 2010).[11] It is important to note, however, that further development of ENLACE may require more in-depth studies, particularly in light of current curricular reforms taking place in Mexico, as well as the uses that the ENLACE assessment may be assigned in the near future. Following is a summary of the elements of the assessment that were reviewed, with preliminary conclusions regarding validity and reliability.

Psychometric model

The tables of specifications for ENLACE were initially established by curriculum experts within SEP who determined the most relevant content to be reflected in each test. The tables were further developed by DGEP staff, with assistance from INEE experts and approved by the Under Secretariat for Basic Education of SEP. For the construction of the assessment, three difficulty levels were established (low, medium and high) to ensure that the tables described the content to be assessed for every grade, difficulty level and subject. The specifications used for construction of test items are publicly available and are open to revisions based on suggestions from teachers, principals and state educational authorities. The tables for mathematics and Spanish, however, have remained unchanged since 2006. The tables allowed for qualitative interpretations of performance differences among students, in order to allow for feedback to improve processes based on results. The scale used to report ENLACE results has a mean of 500 points, with a standard deviation of 100 points, corresponding to 2006 averages as the baseline. ENLACE results have a normal distribution, according to which 99% of students score between 200 and 800 points in each grade and subject. The scale is based on Item Response Theory (IRT) and assigns students to different levels of achievement, with comparability between successive years. Test items are analysed before scoring using a classic model (difficulty index and bi-serial point correlation, as an approximation for item discrimination). The items are then calibrated and the students are classified according to the three-parameter IRT model. A score value is assigned to each student, considering not only the number of correct answers, but also which items were correctly answered. Because a scale is set for each grade and curriculum content-subject, comparisons of scores between different education years or grades is not possible.

Unlike the Rasch model, in the three-parameter model the constructs to be measured are defined before adopting the measurement model. For the development of ENLACE, the existing curriculum guided the construction and selection of items for the test and the parameters were adjusted to the characteristics of the items. In the Rasch model, measurement invariance cannot be obtained simply by using the model on a given set of items; the psychometric model becomes the principle for defining the construct, rather than having the construct guide the development of the test. The three-parameter model used for ENLACE allows the test to reflect the structure of the curriculum without compromising the measurement model. It should be noted that cut scores for achievement levels are not the same in all grades, as they were defined separately for each grade and subject (*i.e.* there is no common scale for all grades).

Dimensionality

The dimensionality of the ENLACE tests is one of the fundamental characteristics that must be analysed in order to determine the structural stability of the results of the assessment over time and hence to allow estimates of actual improvements in learning achieved by students. A recent study of the ENLACE assessment conducted by Lizasoain Hernández and Joaristi Olariaga (2009) concluded that:[12]

- With few exceptions, the ENLACE tests used in the 2008/09 academic year can be considered as essentially one-dimensional or as having weak multidimensionality.
- The results from samples taken from the population and from a control sample do not suggest different dimensional structures.
- The tests can be considered, in general, as having low complexity or a simple structure.
- With regard to possible differences in the dimensionality of the tests, there is some degree of multi-dimensionality in particular grades. This is probably due to the greater complexity of the curriculum content in these grades (*i.e.* in the third year of secondary school).[13]

These findings suggest that the characteristics of ENLACE, combined with the robustness of the test construction models, ensure the correct scaling of students' responses.

Reliability

Reliability of the ENLACE tests is high, based on the internal consistency coefficient (Cronbach's alpha coefficient) calculated. The calculated values for the ENLACE assessment are within the range that is generally accepted as an indicator of highly reliable scores. For the 2006-08 ENLACE applications, for example, the alpha coefficient values vary between 0.75 and 0.92, with averages for Spanish, mathematics and science well above 0.80 (Table 4.2). Based on a comparison of values calculated for the PISA 2000 assessment (OECD, 2002, p. 152, Table 4.1), the reliability of ENLACE is similar and in some cases exceeds that of the PISA 2000 results (unconditioned unidimensional scaling).

Table 4.2

Reliability of ENLACE

	Reliability of ENLACE			
Subject	**Year**	**2006**	**2007**	**2008**
Mathematics	3rd Primary	0.896	0.896	0.912
	4th Primary	0.915	0.898	0.922
	5th Primary	0.896	0.874	0.907
	6th Primary	0.872	0.874	0.910
Spanish	3rd Secondary	0.838	0.789	0.865
	3rd Primary	0.876	0.844	0.879
	4th Primary	0.903	0.900	0.906
	5th Primary	0.809	0.837	0.804
	6th Primary	0.880	0.891	0.910
	3rd Secondary	0.835	0.813	0.752
Science	3rd Primary			0.854
	4th Primary			0.853
	5th Primary			0.818
	6th Primary			0.880
	3rd Secondary			0.804

Source: Zúniga Molina and Gaviria, 2010.

Validity

A study was conducted to assess the concurrent validity of ENLACE in relation to other tests such as the PISA assessment. Since sufficient PISA elements were not publicly available, researchers constructed a special test (SEP-ISA) in agreement with the Australian Council for Educational Research, using test items from the item bank used for the construction of PISA, as well as items previously used for PISA tests and later released. For this study, researchers selected a stratified random sample of 11 717 students in the second and third years of secondary school throughout Mexico. These students took both the ENLACE tests for mathematics and Spanish, and the SEP-ISA test (mathematics and reading comprehension).[14] The correlation of the scales from the different tests, corrected for attenuation, were approximately 0.829 for Spanish/reading comprehension and 0.810 for mathematics (Zúniga Molina and Gaviria, 2010). For comparative purposes, the estimated correlations between scales and subscales of mathematics, reading and science from the PISA 2003 assessment are presented in Table 4.3. The correlations obtained between ENLACE and SEP-ISA are of the same magnitude as those estimated for the subscale problem solving with the other dimensions of mathematics in PISA, ranging from 0.79 to 0.83. Comparatively, these results suggest that the levels of validity of ENLACE are quite high.

Table 4.3
Correlation between subscales of problem solving, reading and science, PISA 2003

	Space and shape	Change and relationships	Uncertainty	Quantity
Space and shape		0.89	0.88	0.89
Change and relationships			0.92	0.92
Uncertainty				0.9
Problem solving	0.79	0.83	0.81	0.82
Reading	0.67	0.73	0.73	0.73
Science	0.73	0.77	0.77	0.76

Source: OECD, 2005a, p. 190, Table 13.4.

Quality of equating process

As described earlier, the specific learning content assessed by ENLACE is determined by the tables of specifications for which four criteria were used to identify, prioritise and focus the curriculum content for the tests: *relevance, plausibility, continuity* and *comprehensiveness.*[15] New versions of the ENLACE tests must be constructed each year given that the test booklets remain in the public domain after application. To ensure that tests are equivalent between consecutive years for the same subject content and grade level, test developers use a variant of the common population design.

The adequacy of the equating process depends on the stability of item parameter estimates and the scores of students: estimates of the parameters should be consistent, regardless of the subset of items used in the estimate. To evaluate the ENLACE assessment in this respect, Zúniga Molina and Gaviria (2010) estimated the item parameters when items were calibrated separately and when these same items were calibrated together with the items from the "pretest form".[16] Two variables corresponding to the difficulty for the item parameter estimates were obtained for each year and subject area. These variables were found to have a very high correlation in all grade levels and subjects, never dropping below 0.993 (Zúniga Molina and Gaviria, 2010, Appendix I). The authors also compared scores obtained by students in one year level (grade), again using two variables for each grade and subject. These values were also found to have a high correlation, never dropping below 0.985 (Zúniga Molina and Gaviria, 2010, Appendix II).

Equating errors were also calculated for ENLACE 2008 and 2009, and then compared with equating error data for the reading comprehension component of PISA 2003. The values for the equating errors were found to be very similar between the 2008 and 2009 ENLACE results, and in some cases lower than equating error values for PISA 2003 reading comprehension. This suggests that the *horizontal* equating process is reliable for ENLACE, allowing for comparisons between student results for each grade and subject over consecutive years (*i.e.* different students, same grade) (Zúniga Molina and Gaviria, 2010, p. 39).

Vertical equating

Although the ENLACE assessment is not designed for vertical equating, SEP and invited experts have conducted feasibility studies to determine the options for further development of ENLACE in the near future to include a vertical scale. This would allow, for example, comparisons of student results between different grade levels (*i.e.* potentially same students, different years). The preliminary studies focused on the results of 104 487 students in the mathematics component of ENLACE in the sixth year of primary school (4 533 classrooms) in Mexico City (presented in Appendix III of Zúniga Molina and Gaviria, 2010). Results from these trials show that it is possible for ENLACE to include a common scale between fifth and sixth grade mathematics. Furthermore, the drop in results for approximately 36% of students between fifth and sixth grades is commensurate with the results of 15-year-old students assessed by PISA. Vertical equating would also allow testing of the cut-off scores defined for each grade level of ENLACE. The cut-off points for all of the years of primary school included in ENLACE (third to sixth), for example, would need to be revised based on a revision of the criteria used to establish them, as well as the validity of the vertical scale. Finally, the degree of the vertical equating errors will also need to be studied before incorporating a common vertical scale in ENLACE. The further development of the ENLACE assessment should incorporate these considerations, as well as others that are outlined in Section 4.3.

Copy factor

As ENLACE is a census assessment (*i.e.* all students in the relevant grade are assessed), supervision and control of test conditions and test application are challenges, particularly given the large diversity of school contexts between and within states. To identify the magnitude of probable answer copying in the applications, two different methods are used. Although the average percentages of probable cheating reached a high of 7.0% in 2008 compared to 4.5% in 2006, the trend decreased in 2009 with an average of 6.5% (Table 4.4). Initial results from the 2010 application of ENLACE confirm that answer copying has not continued to increase. Furthermore, a consideration of the general effects of copying conducted by Zúniga Molina and Gaviria (2010) shows that even for 2008, the estimates of validity and reliability for ENLACE remain largely unaffected.

Table 4.4

Percentages of probable test cheating cases detected for ENLACE 2006 to 2009

Year	2006	2007	2008	2009
3rd Primary	6.28	5.50	10.16	6.97
4th Primary	4.90	5.60	7.53	6.65
5th Primary	4.14	3.68	4.78	4.45
6th Primary	4.10	4.86	5.48	4.75
1st Secondary				1.54
2nd Secondary				3.79
3rd Secondary	3.24	3.14	7.08	6.54

Source: OECD, 2005a, p. 190, Table 13.4.

Further monitoring and analysis of the answer copy factor should continue, however, and measures to address this should be considered for every application. Additional resources and supervisory mechanisms should also be included in the planning of the ENLACE assessment.

4.3 CHALLENGES AND OPPORTUNITIES FOR FURTHER DEVELOPMENT OF THE ENLACE ASSESSMENT SYSTEM

In Mexico as in other better-performing educational systems, consensus is emerging on the benefits of clearly defining the progress students are expected to achieve in the acquisition of skills and competencies. In recent years, therefore, SEP has undertaken curricular reforms focused on the development of student skills, with a major emphasis on achieving specific learning outcomes. These reforms have advanced significantly in the pre-school and secondary levels, and are in an experimental phase for primary school. To the extent that these reforms focus on developing competencies and skills for life, ENLACE will need to reflect these changes. With a standards-based framework in Mexico, curriculum-referenced testing will need to evolve to reflect standards of competencies and skills that may be established. Clearly defined content and performance standards for students, developed as part of the curricular reforms in Mexico, could serve as the anchor and reference for teacher planning, educational materials, teaching practices, capacity-building and professional development, and ultimately assessments. Perhaps most importantly, clear standards for student learning and growth would also provide a coherent view and create shared *expectations* that all educational levels and actors can share and work towards, including state educational authorities, parents and school-level committees and councils. In this context, the following points should be considered:

- The need to preserve the levels of reliability, validity and structural simplicity and stability already achieved by ENLACE while taking into account the curricular transformations being considered and that may be implemented.
- The need to establish a clear development programme for ENLACE, defining policy objectives, targets and timeframes. The further development of ENLACE should include relevant studies to determine the vertical comparability for students and groups, in order to measure progress towards defined learning expectations. Measures of student progress should be net of socio-economic and other relevant factors, to identify the contributions of schools, school zones, regions and states towards student outcomes.
- The need to address administrative, technical and logistical considerations, in order to allow for more reliable measures of student growth over time towards specific learning objectives (*e.g.* as defined in content standards of student learning). Addressing these elements would not only make the ENLACE assessment more robust but would also contribute to strengthening the evaluation framework in Mexico, for accountability and for improvement efforts (Chapter 3). Furthermore, addressing the following items would permit the development of value-added methods that are presented in more detail in Chapter 5 of this report and in its sister publication.

Administrative conditions

- Completeness and consistency in the identification of students, teachers, schools and principals where appropriate. There must be specific mechanisms to detect and incorporate individuals who might not be in the databases and to correct inconsistencies. Student and teacher mobility within and across school zones and states should be identifiable with proper tracking.
- Unified individual dossiers for students to accompany each student throughout his or her entire school life. Information on the results of assessments, including ENLACE, must be included in order to determine progress in learning. This dossier could also be used for at-risk students and for efforts to reduce drop-out rates.
- Uniform and unique references for cities, towns, municipalities and schools.
- Capacity to link and match the achievement of students on ENLACE with the teachers who have taught them.
- Capacity to match each item in the database of students with their counterparts in the databases of teachers, principals and schools, in order to determine the contribution of different teachers, or the entire school, to the learning gains of each student.

Logistical conditions

- Substantial improvement in the conditions controlling the application of the test. This includes mechanisms to limit the potential for undesired behaviours from teachers and principals as well as addressing the issue of answer copying. The need for security, supervision, and adequate and standardised conditions for the application of ENLACE will only increase and adequate resources should be considered by SEP and state educational authorities.

Technical conditions

- Continued technical robustness of ENLACE assessments. The demonstrated validity, reliability and internal consistency of ENLACE suggest that only substantial changes in the curriculum would warrant a corresponding transformation of ENLACE. In light of the curricular reforms being considered and implemented in Mexico, however, these may offer an opportunity to plan the further development of ENLACE to ensure alignment, coherence, cognitive demand and breadth that are commensurate with policy objectives.

- Implementation of a vertical alignment design for all grades and content subjects that allows the calculation of educational progress for each student. SEP and invited researchers will need to calculate comparability errors and verify the extent to which the errors on the same cohort, in successive grade levels, are additive and whether the magnitude of these cumulative errors prevents the scale going beyond two or three year levels, for example. The cut-off points should be redefined so that they are consistent across consecutive grades.

4.4 SUMMARY RECOMMENDATIONS FOR MEXICO

Based on the considerations presented earlier, the following are the main summary recommendations for Mexico regarding the importance of assessing student learning outcomes, the opportunities afforded by the ENLACE assessment, and the challenges and opportunities for its further development:

- ***Student learning and growth as the basis of accountability and standards requires multiple, cross-referenced, valid and reliable measures.*** Because all of the current measures and instruments of student learning and growth (standardised tests, teacher assessments, portfolios of student work and observation, among others) present potential sources of error and bias, a complementary approach that uses valid evidence from multiple sources should be gradually developed to assess current instruments in Mexico, estimate costs, and determine the capacity-building and instrument development that are required. With clear content and performance standards of what students are expected to know and know how to do, for example, measures that reflect the learning and growth expected from students can be further developed.

- ***The use of student performance data should be accompanied, when possible, with complementary and reliable measures of student learning, as these are developed, tested and validated.*** The relative importance of student data and school-based or teacher assessments can be redefined as needed by the policy objectives and consequences resulting from the assessments. Australia, Alberta (Canada) and Hong Kong-China are examples of better-performing systems that attempt to combine standardised assessments with school-based assessments (*e.g.* locally graded but externally moderated), student projects, and extended papers.

- ***Student performance data, such as those from the annual ENLACE assessment in Mexico, can play an important role in accountability and school improvement efforts.*** Current efforts by SEP and state educational authorities regarding the presentation and use of ENLACE demonstrate the high degree of social acceptance and potential of ENLACE. Student performance data aggregated at the group, school, zone or state levels can be employed in static, improvement, or growth models, depending on the specific purpose of the policy levers and programmes in Mexico.

- ***A specific development programme should be established for the ENLACE assessment, considering issues of cognitive demand, curricular alignment and coherence.*** The best-available evidence on student learning progression and standards should be considered. The development of ENLACE should set clear stages and goals that address technical (*e.g.* vertical equating), administrative (*e.g.* unique student, teacher and school identifiers and linkages) and logistical (*e.g.* improved test supervision) considerations.[17] With expanded use of the ENLACE assessment in the future, enhanced supervision and security of test administration, for example, should beaddressed. The programme should also have a long-term vision that takes internationally benchmarked content and performance standards into account. As content and performance standards are established in Mexico, student performance data can be used, in conjunction with analytical models (*e.g.* growth) for specific policy objectives and programmes. Throughout the process, consideration should be given to the alignment and coherence between standards, assessment and professional development for teachers. A clear vision of the evaluation framework in Mexico should allow for the distinct but complementary purposes of different assessments (*i.e.* ENLACE, EXCALE, or possible school-based assessments), and how they should continue to develop in the future within a common national framework.

- ***With student performance data and appropriate growth models, low performers, high performers and cases needing follow-up observation can be identified.*** As the assessment and evaluation process becomes more established, consequences such as incentives, further observation, and assistance to schools and teachers can be linked to the results. This implies both a gradual development of the process and the possibility of having multi-stage consequences and responses to the results.

Notes

[1] For reference, the average performance of students in Alberta in PISA 2006 was significantly above the Canadian average, which was already among the top performers, along with Hong Kong-China (OECD, 2007; Bussiere, Knighton and Pennock, 2007). Australia was among the top-10 performing economies, out of 57 (OECD, 2007).

[2] The most commonly tested subjects are mathematics and the national language, with science included in only seven out of 29 countries for which information was available (OECD, 2008).

[3] With constructed response items, once the question is stated, the load of work falls on the marker's shoulders, and the correspondence between the score assigned to a particular student and her/his cognitive status depends on the marker's dexterity in correctly detecting the telling signs of that cognitive status. With multiple-choice items, a considerable amount of work for the assessment is conducted previously, when dividing the cognitive task into the relevant steps where the different cognitive levels must be identified through the different combinations of the alternatives.

4. For the government of Alberta, Canada, teachers' professional responsibilities should include the marking of provincial achievement tests *(http://education.alberta.ca/department/ipr/commission/report/reality/governance/bargmodel.aspx)*, while marking of student assessments "need not be done by teachers" in a document presented by the largest union in the United Kingdom, the NASUWT (cited in Stevenson, 2004, p. 233).

5. For example, based on the ICT Development Index that combines access, use and skills data from the International Telecommunication Union, Korea ranked 2nd overall, Finland 9th while Mexico placed 75th, below Chile (48th), Turkey (59th) and Brazil (60th) (ITU, 2009).

6. A general resource regarding testing is provided by the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association and National Council on Measurement in Education, 1999).

7. This section is largely based on expert contributions commissioned by the OECD as part of the Co-operation Agreement with the government of Mexico. The two working papers are *Challenges and Opportunities for the Further Development of the ENLACE Assessment for Evaluation and Teacher Incentives in Mexico* (Zúniga Molina and Gaviria, 2010), and *State-Level Teacher Evaluation and Incentive Practices in Mexico: Diagnostic Study* (Salieri *et al.*, 2010).

8. Administered by the National Institute for the Evaluation of Education (*Instituto Nacional para la Evaluación de la Educación*, INEE), these are the Educational Quality and Achievement Exams (*Exámenes de la Calidad y el Logro Educativos*, EXCALE). EXCALE exams are sample-based assessments administered in four-year cycles to students of certain key grade levels at the pre-primary, primary and secondary levels. The assessment in 2011 will be for third-year pre-primary students in Spanish and mathematics, followed by an exam on Spanish, mathematics, natural sciences and social sciences for third-year secondary students in 2012.

9. Schools are classified according to the different basic education programmes in the country. The classifications are used to compare results of schools that have similar characteristics in terms of the student population and the resources available for students and teachers.

10. The application of ENLACE to the control sample used for equating purposes is accompanied by a context questionnaire. Information related to the characteristics of the school is taken from the stratification variables used for the sampling process. Although a few studies relating these contextual variables to ENLACE performance have been undertaken, no further operational use of this information has been made.

11. Considerations of consequential validity are not included in this assessment given the multiple uses and breadth of purpose to which ENLACE is currently subjected.

12. The original study is in Spanish and is included as an annex to the Zúniga Molina and Gaviria (2010) paper prepared for the OECD on which this section is based.

13. Until 2009, the third year ENLACE tests were designed to reflect the cumulative content of the secondary level as a whole (*i.e.* tests included content for the first and second years as well). This may explain the findings of Lizasoain Hernández and Joaristi Olariaga (2009).

14. The correlation values between latent variables of ENLACE and SEP-ISA are included in Zúniga Molina.

15. *Relevance* refers to the relative importance attributed by experts to each topic, and depends on the depth of treatment of the different subjects in textbooks; *plausibility* refers to the feasibility of developing multiple-choice items in relation to the content subjects to be included in each particular learning test; *continuity* refers to the extent to which specific content is part of a teaching sequence that extends beyond a particular year-grade; and *comprehensiveness* refers to the level of inclusion of other content associated with lower degrees of complexity (Zúniga Molina and Gaviria, 2010).

16. As ENLACE is applied annually, every year a parallel test (referred to as the "pre-test") is developed and applied to a control sample of students who also take the normal test given to students that year (this is referred to as the "operational form" of ENLACE). The parallel test items are calibrated in conjunction with the operational form of the test and are then used to form the new test for the following year.

17. The specific technical, administrative and logistical recommendations on further development of the ENLACE assessment are presented in Chapter 5.

References

American Council on Education (Linn, R.L., ed., 1993), *Educational Measurement*, Oryx Press, AZ.

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

Baker, E. (2003), *Multiple Measures: Toward Tiered Systems*, University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles.

Baker, E. (2004), *Aligning Curriculum, Standards, and Assessments: Fulfilling the Promise of School Reform*, University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles.

Baker, E. (2010), "Assessment and Accountability", expert paper commissioned by the OECD for the Co-operation Agreement between the OECD and the government of Mexico.

Burstein, J.C. (2003), "The e-Rater Scoring Engine: Automated Essay Scoring with Natural Language Processing", in M.D. Shermis and J. Burstein (eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective*, Erlbaum, Mahwah, NJ, pp. 113-122.

Chung, G.K.W.K. *et al.* (2001), "Knowledge Mapper Authoring System Prototype" (final deliverable to OERI), University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles.

Chung, G.K.W.K. and **E.L. Baker** (2003), "Issues in the Reliability and Validity of Automated Scoring of Constructed Responses", in M.D. Shermis and J. Burstein (eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective*, Lawrence Erlbaum, Mahwah, NJ, pp. 23-40.

Lizasoain Hernández, L. and **L. Joaristi Olariaga** (2009), "Estudio de la dimensionalidad de las pruebas ENLACE (2008) mediante técnicas factorials clásicas y métodos no paramétricos basados en TRI – Informe Preliminar", included as Technical Annex V to the Zúniga Molina and Gaviria (2010) expert paper commissioned by the OECD for the Co-operation Agreement between the OECD and the government of Mexico.

Government of Alberta (Canada) – Education (2009), "The Alberta Student Assessment Study: Final Report", Edmonton, Alberta.

Instituto Nacional para la Evaluación de la Educación (INEE) (2010), *Panorama Educativo de México: Indicadores del Sistema Educativo Nacional 2009 Educación Básica*, INEE, Mexico City.

International Telecommunication Union (ITU) (2009), "Measuring the Information Society: The ICT Development Index", ITU, Geneva.

Moriconi, G.M. (2009), "The Development Index of Basic Education and Teacher Evaluation in Brazil", Presentation given at the OECD/SEP International Workshop *"Towards a Teacher Evaluation Framework in Mexico: International Practices, Criteria, and Mechanisms"*, 1-2 December 2009, Mexico City.

Organisation for Economic Co-operation and Development (OECD) (2002), *PISA 2000 Technical Report*, OECD Publishing, Paris.

OECD (2005a), *PISA 2003 Technical Report*, OECD Publishing, Paris.

OECD (2005b), *Formative Assessment: Improving Learning in Secondary Classrooms*, OECD Publishing, Paris.

OECD (2007), *PISA 2006: Science Competencies for Tomorrow's World*, OECD Publishing, Paris.

OECD (2008), *Education at a Glance 2008: OECD Indicators*, OECD Publishing, Paris.

OECD (2009), *Assessment and Innovation in Education*, OECD Working Paper No. 24, OECD Publishing, Paris.

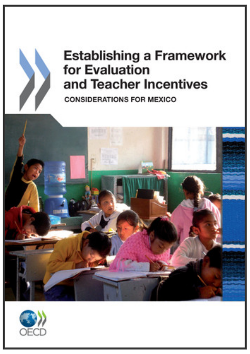
OECD (2010), *La medición del aprendizaje de los alumnos: Mejores prácticas para evaluar el valor agregado de las escuelas*, OECD Publishing, Paris.

Parandekar, S.D., E. Amorim and **A. Welsh (2008),** "Prova Brasil – Building a Framework to Promote Learning Outcomes", Note No. 121, The World Bank, Washington, DC.

Salieri, G., L. Santibañez and **B. Naranjo** (2010), *State-Level Teacher Evaluation and Incentive Practices in Mexico: Diagnostic Study*, study commissioned by the OECD for the Co-operation Agreement between the OECD and the government of Mexico.

Stevenson, H. (2007), "Restructuring Teachers' Work and Trade Union Responses in England: Bargaining for Change?", *American Educational Research Journal*, Vol. 44(2), pp. 224-251.

Zúniga Molina, L. and **J.L. Gaviria** (2010), *Challenges and Opportunities for the Further Development of the ENLACE Assessment for Evaluation and Teacher Incentives in Mexico*, expert paper commissioned by the OECD for the Co-operation Agreement between the OECD and the government of Mexico.



From:

Establishing a Framework for Evaluation and Teacher Incentives

Considerations for Mexico

Access the complete publication at:

https://doi.org/10.1787/9789264094406-en

Please cite this chapter as:

OECD (2011), “Using Student Learning Outcomes to Measure Improvement”, in *Establishing a Framework for Evaluation and Teacher Incentives: Considerations for Mexico*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/9789264094406-7-en