Chapter 3
How to Ensure the Validity of National Learning Assessments? Priority Criteria for Latin America and the Caribbean

María José Ramírez and Gilbert A. Valverde

3.1 Introduction

The number of learning assessments in Latin America and the Caribbean (LAC) has grown significantly. Countries introduce these assessment systems to monitor how well their educational systems pursue curricular objectives and to foster improvements in the system in general, and in student learning in particular. It is reasonable to ask, therefore, how helpful these tests are in substantiating inferences about the achievement compared to goals proposed in curricular policies, to what extent their results can be interpreted as a reflection of such learning, and whether they can be effectively used to promote improvement in learning. To answer these questions, validity evidence is needed.

So far, Latin American countries have been more concerned with installing the assessments than with validating them. While most countries introduced national assessments in the 1990s, these assessment regimes have been unstable. In weak institutional contexts, most resources are often put into installing the assessments. There is little capacity to document validity evidence (e.g., in technical reports) and

[1] For the purposes of this study, we accept these stated objectives as the fundamentals of the assessment policy in the region, although there may and should be other stated and undeclared objectives for these assessments.

We appreciate Elisa de Padua's valuable collaboration in collecting and analyzing information on validation practices in learning assessment programs. We also thank all the professionals of the assessment programs contacted. Finally, our thanks to Patricia Arregui (GRADE) for her valuable contributions and comments.

M. J. Ramírez
Independent Consultant, Education, Alexandria, VA, USA

G. A. Valverde (✉)
University At Albany, State University of New York, Albany, USA
e-mail: gvalverde@albany.edu

J. Manzi et al. (eds.), *Validity of Educational Assessments in Chile and Latin America*,
https://doi.org/10.1007/978-3-030-78390-7_3

even less to conduct validity studies or external audits. In such contexts, a focus on the validity can be viewed as a threat to the legitimacy of the assessments—a legitimacy that was so difficult to build, in the first place.

However, now that assessments in the region have matured and become an indispensable part of public discussion on educational issues, it is imperative to prioritize the validity agenda. That is, an agenda that ensures that evidence is collected to support the interpretations, uses, and policy decisions associated with the assessments.

Assessment validation is a must: it is necessary to ensure their political and technical feasibility. Without validity evidence, how could one know whether assessments measure what they are supposed to measure, or whether their results truthfully reflect student learning? Without validity evidence, it is impossible to know whether assessments really help improve the educational system.

Evidence of validity is the basis for judging the quality of the information used to make decisions. Making inferences from information of questionable quality is no better than making decisions without any information at all. It is like measuring temperature without any idea whether the thermometer is working properly: the risks of misdiagnosis would be extremely high—and who knows if the proposed remedies would help? Not having evidence of validity can lead to what policy analysts call "type 3 error": solving the wrong problem (Mitroff & Featheringham, 1974).

The cost of introducing a national assessment without evidence of validity is too high: it is equivalent to making an investment with no way of judging its value. One can see how important it is to validate an assessment if we consider the costs of not doing it. Consider, for example, the political costs of reporting lowered educational achievement in a country when student results have, in fact, improved—or the social costs of incorrectly classifying a school as "insufficient" or "underperforming." Lack of validation evidence may lead to criticism that could damage or bring down national assessments.

The international community has developed various technical quality standards for validating learning assessments (see American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014; UNESCO Institute for Statistics [UIS] & Australian Council for Educational Research [ACER], 2017; Darling-Hammond et al., 2013). However, these standards have mainly been produced for developed countries with their higher (compared to LAC) average levels of student achievement and typically more sophisticated assessment systems.

This chapter proposes ten priority criteria or quality standards for the validation of learning assessments in LAC. These have been classified into three dimensions or sources of evidence: (1) the dimension related to test alignment with the official curriculum, (2) the dimension related to the curricular validity of performance levels used to report the results of assessments, and (3) the dimension of consequential validity: the assessment's impact on the improvement of the education system in general, and on the improvement of learning in particular.

The dimensions and criteria selected were chosen based on: (1) learning assessment standards commonly accepted in the international community, (2) information collected by the authors in 2016 about validation practices in learning assessments in nearly 50 countries, including those representing best practices, and (3) the authors' 20 years of academic and professional experience in designing and implementing assessments, consulting, conducting external audits, and providing international technical assistance in learning assessment in LAC and other regions of the world.

This chapter addresses the validation of assessments from a conceptual perspective, and delineates the situation observed in LAC with respect to the different assessment dimensions examined. It also presents examples of good validation practices from Canada and the United States. The arguments presented in this introduction are further developed in the chapter by Valverde and Ramírez where the authors present in-depth case studies on validation practices in different Latin American countries.

3.2 What Does It Mean to Validate Assessments?

We define learning assessments as national tools for measuring the progression towards achieving curricular objectives (competencies, content, or skills, as appropriate). Their main purpose is to monitor learning at the country level, and to promote the improvement of the education system in general and of learning in particular.[1] To this end, standardized tests are administered to all students (census), or to nationally representative samples, at key points in their school trajectories (e.g., at the end of the first or second cycle of primary education). The tests are usually paper-and-pencil[2] and include multiple-choice questions, problems, or items and, to a lesser extent, open-ended questions. Assessment results are typically used to inform educational policy and practice, and may have consequences associated with them (e.g., incentives).

Validating an assessment means gathering evidence to support its interpretations, uses, and expected consequences. For example, if the assessment measures the attained curriculum, there should exist documentation (evidence) of the alignment between the tests and the curriculum. If the assessment says that teachers will use the results to plan and improve their lessons, there must be evidence to show this is possible. The definition proposed here is based on a unitary concept of validity which also includes a dimension of consequential validity, or the impact of assessments on the education system.

[2] Although computer tests are becoming more common.

3.3 Validation Criteria

Table 3.1 presents ten priority criteria for the validation of learning assessments in LAC. Each criterion is accompanied by an explanation and examples of evidence to validate the interpretation, uses, and expected impacts of the assessments. A distinction is made between evidence of products (e.g., assessment instruments) and evidence of processes (methods and procedures for producing the product).

Table 3.2 presents a checklist for reviewing the validity of national assessments. It operationalizes the criteria of the priority dimensions, presenting more specific examples of evidence for validation. Each row or type of evidence refers to one or more priority criteria identified in Table 3.1. The table also provides space for making notes on each type of evidence (e.g., "Alignment review panels are created, but there is no documentation").

Finally, we want to emphasize that while in any assessment there are other key standards or validity criteria, they are not the subject of this chapter (e.g., those related to field operations or data processing). Since in LAC, assessments are predominantly used to monitor or verify compliance with learning goals established in curricular policies, we consider validation with respect to the curriculum to be a priority. Hence, the following criteria only refer to the dimensions identified as priorities: the dimension of alignment of tests with the official curriculum, the dimension of curricular validity of performance levels, and the dimension of consequential validity of assessments.

3.4 Dimension of Test Alignment with the Official Curriculum

In LAC, it is a priority to ensure that curricular learning assessment systems collect validity evidence that their tests effectively measure the curriculum. That is, they need to be aligned with the competencies, capabilities, content, or other equivalents described there. Table 3.1, in the dimension of test alignment with the official curriculum, presents three priority criteria of validity for the region, along with evidence necessary to validate each criterion. These are: (1) The design of the assessment is justified in reference to the curriculum, (2) The assessment domain is operationalized with actual student learning in mind, and (3) Test results allow accurate and unbiased monitoring of the achievement of curricular learning over time. Table 3.2 presents a more detailed checklist with examples of evidence for the validation of alignment between the curriculum and the tests.

Table 3.1 Priority dimensions and criteria for assessing the validity of national learning assessments in Latin America and the Caribbean

Criterion		Explanation	Validity evidence
Dimension of test alignment with the official curriculum			
1	The design of the assessment is justified in reference to the curriculum	There is an explanation of the why, whom, what, how, and when of the assessment, with explicit references to the curriculum. Possible interpretations and intended uses of the results of the assessments are identified. The assessment domain is specified (e.g., mathematics, language) referring to the competencies, objectives, content, or skills, and performance levels defined in the curriculum	• Products: (a) general design of the assessment (the why, whom, what, how, and when it is evaluated) with references to the curriculum, (b) test specifications, (c) final version of the test • Processes: methods and procedures for: (a) the designing of the assessment, (b) developing test specifications, and (c) producing the final version of the test

(continued)

Table 3.1 (continued)

Criterion		Explanation	Validity evidence
2	The assessment domain is operationalized with actual student learning in mind	Recognizes the need for assessment to be aligned with what all students know and can do, from the most to the least advanced. The assessment domain and test specifications cover a wide spectrum of learning. The assessment results allow a characterization of what all students know and can do, even those who have more learning difficulties	• Products: (a) final version of the test, (b) assessment domain and test specifications, (c) test results, and (d) information on actual student learning (e.g., results of classroom or international assessments) • Processes: methods and procedures for (a) operationalizing the assessment domain and making specifications of tests, (b) collecting information about the real student learning (e.g., classroom observations, reviewing the results of classroom and international assessments)

(continued)

Table 3.1 (continued)

Criterion		Explanation	Validity evidence
3	Test results allow accurate and unbiased monitoring of the achievement of curricular learning over time	Test results can be interpreted in terms of achieving curricular objectives. The results are published indicating the margin of error or uncertainty associated with the scores, percentages, differences, or other statistics. The margin of error is within the acceptable range, given the consequences associated with the assessment. The results are free of bias (e.g., gender, cultural, linguistic, ethnic) The results of different tests have been equated and put on the same score scale	• Products: achievement scores and measurement errors, equating, and psychometric analysis (e.g., difficulty, discrimination, differential item behavior) • Processes: methods and procedures for calculating learning outcomes and measurement errors, applying equating, and calculating psychometric characteristics of the tests
Dimension of curricular validity of the performance levels			
4	Performance levels are aligned with the curriculum	Performance levels are described and justified by different aspects of the curriculum (e.g., competencies, complexity, contexts). Performance levels represent necessary learning milestones towards achieving curricular goals. One of the performance levels corresponds to the achievement of the curricular objectives	• Products: final version of the curricular specifications and psychometric specifications of the performance levels • Processes: methods and procedures for developing performance levels aligned with the curriculum

(continued)

Table 3.1 (continued)

Criterion		Explanation	Validity evidence
5	Performance levels are operationalized with actual student learning in mind	Performance levels are designed to ensure that the assessment can describe what all students know and can do. They cover a wide spectrum of learning, from the most advanced to the most basic. There is a performance level that describes what the lowest performing students know and can do	• Products: (a) results by performance levels, (b) final version of the curricular specifications and psychometric specifications of performance levels • Processes: methods and procedures for: (a) collecting information on real student learning (e.g., classroom observations, classroom and international assessments), and (b) using such information to develop performance standards

(continued)

Table 3.1 (continued)

Criterion		Explanation	Validity evidence
6	Performance levels describe qualitatively different stages of learning	Performance levels define milestones on a learning trajectory. What students know and can do at one level is clearly different from what they know and can do at the next. Students are classified into performance levels according to whether they have achieved the level of mastery associated with the appropriate cut scores. There is one performance level that is defined by default and that corresponds to students who do not reach the lowest cut score on the scaled score	• Products: (a) final version of the curricular specifications and psychometric specifications of performance levels, with cut scores, descriptions of what students know and can do at each level, and titles for each level; (b) results according to performance levels • Processes: methods and procedures for (a) developing performance levels and (b) sorting students into different levels of performance

(continued)

Table 3.1 (continued)

Criterion		Explanation	Validity evidence
7	Performance levels balance stability and change, in the context of a dynamic curriculum policy	Two versions of performance levels are used: 1. Updated performance levels to monitor changes in learning in the short term (e.g., from one year to the next). They are reviewed and updated according to curricular reforms and changes 2. Invariant performance levels to monitor changes in learning in the long term (e.g., 10- or 20-year trends). They describe core learnings that are usually less affected by curricular adjustments	• Products: (a) updated performance levels, (b) learning outcomes associated with the updated performance levels, (c) invariant performance levels, (d) learning outcomes associated with invariant performance levels • Processes: methods and procedures for: (a) reviewing updated performance levels, (b) assess students using the updated performance levels, (c) developing invariant performance levels, (d) assessing students using the invariant performance levels
Dimension of consequential validity of the assessments			

(continued)

Table 3.1 (continued)

Criterion		Explanation	Validity evidence
8	Assessment results are effectively communicated	Results are clearly, correctly, and in a timely manner disseminated in different media and formats (e.g., reports, workshops, brochures, videos), suitable for different audiences	• Products: (a) communication products used for dissemination of results and associated information on the population of students tested and their schools and communities, (b) communication strategy used (e.g., communication products, audiences, timing), (c) levels of access, public trust, and understanding of results by different stakeholders • Processes: methods and procedures for designing the communication strategy (e.g., consultation with key audiences or stakeholders)
9	There are formal mechanisms to support the use of assessments to improve learning	There are formal/institutional mechanisms that encourage the use of assessments for improvement (e.g., school improvement plans, guidelines for school supervision, guidelines for teacher training, pedagogical resources consistent with the curriculum and the actual learning levels of students)	• Products: (a) an inventory of formal mechanisms that use assessment inputs • Processes: methodology and procedures for collecting information on the uses of assessment results by different educational actors

(continued)

Table 3.1 (continued)

Criterion		Explanation	Validity evidence
10	There are formal mechanisms to monitor the impact of assessments on the education system	Studies on the intended and unexpected consequences of assessments in the education system are regularly carried out. E.g., evaluations of how assessments have impacted student learning, teacher practice, or parental involvement, or how they are used to inform policy	• Products: studies on the consequences of the assessments • Processes: formal mechanisms to carry out studies of consequential validity: commissions or audits, internal or external financing for such studies

Note Table constructed by the authors

Table 3.2 Checklist for Assessing the Validity of National Learning Assessments in Latin America and the Caribbean

Criterion	Validity evidence	Comments on validity: (a) of the process of developing the test instrument (b) of the test instrument
Dimension of test alignment with the official curriculum		
1	The purposes of the assessment ("what for?") are formulated with explicit reference to the curriculum	
1	Possible interpretations and intended uses of the assessment results are identified	
1	The overall design of the assessment is described: who is assessed; what, how, and when are measurements carried out, with explicit reference to the curriculum	
1	The assessment domain is specified (e.g., mathematics, language) with reference to the competencies, objectives, contents, or skills defined in the curriculum	
1	The quantity and distribution of items in different categories is justified (e.g., contents, capacities) based on the curriculum	
1	The assessment domain is specified (e.g., mathematics, language) and performance levels are considered in the specification	
2	There is an explanation specifying the empirical relationship between the definitions of the assessment domains and the actual learning level of students who are lagging	
2	There is an explanation specifying the empirical relationship between the test specifications (or blueprint) and the learning of students who are lagging	
3	There is a clear explanation showing how test items constitute an adequate sample of the assessment domain	
3	The test items are correctly classified according to the categories of the test specifications	
3	The actual number and distribution of items in different categories matches the intended test specifications	
3	Test specifications and performance levels are used to guide item development	
3	Test results are published indicating that there is a margin of error or uncertainty associated with the scores, percentages, differences, or other statistics	
3	The level of reliability (accuracy) of the tests is reported, which should be equal to or greater than Cronbach's alpha = 0.70	
3	The classification error related to performance levels is reported, and its size is appropriate to the severity of the consequences associated with the tests	
3	Tests and items were subjected to Differential Item Functioning (DIF) analyses to avoid bias in the scores	

(continued)

Table 3.2 (continued)

Criterion	Validity evidence	Comments on validity: (a) of the process of developing the test instrument (b) of the test instrument
3	There is a clear explanation of how different test booklets administered in a single year were put on the same scale of scores through *equating*	
3	There is a clear explanation of how different tests administered in different years were put on the same scale of scores through *equating*	
Dimension of curricular validity of the performance levels		
4	Performance levels are described and justified by referring to different aspects of the curriculum (e.g., competencies, complexity, contexts)	
4	The level of performance that corresponds to the achievement of the curricular objectives for the grade or cycle assessed is indicated	
5	The relationship between the lowest performance level and what the students with the weakest learning levels know and can do is shown	
6	What students know and can do at one performance level is qualitatively different from what they know and can do at the next performance level	
6	The distance between the cut scores associated with the performance levels is at least half of a standard deviation (SD = 0.50) on the score scale	
6	How students are classified into different performance levels depending on whether or not they reach the score associated with each level is explained	
6	The results by performance level distinguish between students that reach or don´t reach the lower cut score. Students who do not reach it are classified at a level that is defined by default	
6	The lowest cut score is set so that no more than 25% of the students fall below it	
7	There is a clear explanation of how performance levels are reviewed and updated following curricular reforms and adjustments	
7	The use of updated performance levels to monitor learning in the short term is described	
7	The use of invariant performance levels to monitor learning over the long term is described	
Dimension of consequential validity of the assessments		

(continued)

Table 3.2 (continued)

Criterion	Validity evidence	Comments on validity: (a) of the process of developing the test instrument (b) of the test instrument
8	Assessment results are published in time to inform decisions made by agencies and stakeholders in the education system, according to the objectives of the assessments	
8	Assessment results are disseminated in different media and formats (e.g., reports, workshops, brochures, videos) for different audiences (e.g., teachers and school leaders, parents, general public)	
8	The communication plan used to disseminate assessment results and information is described	
8–9	The extent to which educational stakeholders (e.g., parents, teachers, teacher educators, and policymakers) value, understand and use assessment results and information to make decisions about educational practice or policy is documented	
9	There are a variety of formal or institutionalized mechanisms for using assessment results and information	
10	There are formal mechanisms for regularly collecting information on the expected and unintended consequences of assessments	

Note Prepared by the authors

3.4.1 Criterion 1: The Design of the Assessment is Justified in Reference to the Curriculum

To fit their purpose of monitoring curricular learning achievement, assessments must be aligned with the official curriculum. That is, the tests must measure the objectives, competencies, content, or skills (as appropriate) set out in the curriculum. Alignment between tests and curricula is essential so that assessment results can be interpreted as the achievement of curriculum objectives, and in particular, so that a higher score can actually be interpreted as an indicator of higher levels of mastery of the curriculum than a lower one. Alignment with the curriculum also means that performance levels correctly identify students who are at different stages of learning.

The assessment should be accompanied by documentation evidence to judge the degree of alignment between the current curriculum and the tests. The assessment framework should clearly indicate the purposes and expected uses of the assessments based on the curriculum; for example, whether the purpose of the assessment is to monitor the achievement of curriculum objectives at the national, sub-national (e.g., regions), school, or classroom level. The framework should provide guidelines on how to correctly interpret assessment results in terms of the curriculum and use them to improve learning.

The design of the assessment should be based on the current curriculum. Test specifications describe and justify the objectives, competencies, content, or capabilities (as appropriate) to be assessed; test format (e.g., paper-and-pencil or computer-based) and type of items (e.g., multiple-choice or open-ended questions); situations or contexts in which students need to demonstrate what they have learned (e.g., abstract or applied mathematics problems); cognitive complexity of tasks to be performed; times to complete test booklets; and so on. Test specifications usually include double-entry tables indicating the number and type of items to be included in each test, classified according to different categories; for example, a table with axes of contents and skills to be evaluated in mathematics, with the number of items to be included in each cell (crossing contents and skills). The distribution of items in different categories is an indication of the importance assigned to each one in the assessment (weights). In LAC, validity evidence usually focuses on test specifications. Often, this is the only type of documentation available on instrument design.

Indeed, the purposes of assessments are often described in very general terms in LAC, without specifying the type of interpretations they are designed to make possible or their appropriate uses. Specifying the purposes of the assessments is more difficult in the context of curricula that are not written with the express purpose of being measured (see chapter by Valverde and Ramírez in this book).

Performance levels should inform the development of test items. This is fundamental for measuring each level well, with items that point to the competencies, curricular content, or capabilities that characterize them. Guidance for item development can come from either preliminary or definitive descriptions of performance levels. Preliminary descriptions can be used when there is no empirical evidence of student performance yet (i.e., before the tests are administered). Definitive descriptions can be used after the performance levels have been developed and adjusted in relation to the empirical evidence, i.e., to the test results.

In LAC, performance levels are not systematically used to develop new items. This often occurs because performance levels are not part of the curriculum documents. They are usually created afterwards, once test results are available. Given that the performance levels seek to describe what students know and can do in relation to the curricular objectives, it is desirable that these be developed, at least in a preliminary way, before designing the tests, and used as input for the test specifications. This typical absence of specification of performance levels in the design of the test distinguishes assessment programs in LAC from the more sophisticated assessment programs in the world.

3.4.2 Criterion 2: The Assessment Domain is Operationalized by Taking into Consideration Actual Student Learning

Test design should take into consideration actual student learning. This ensures that an assessment can monitor the learning levels of all students, from the most advanced

to the ones that are lagging behind. Therefore, test specifications and performance levels must account for the competencies, knowledge, and skills of all, including the least advanced students. Consequently, items should be developed so that they cover the full range of students' abilities.

There is a clear tension between measuring curricular objectives and measuring what students who are lagging behind really know and can do. Curricular objectives usually correspond to the most sophisticated and difficult things measured in the tests. In LAC, what struggling students know and can do is often not measured, as it is considered too easy, basic, and distant from the curricular objectives. This does not allow for the visibility of these students, who are precisely those who need more support to improve their learning. There is a close correlation between learning and students' socio-economic background, repeatedly confirmed by research and assessments in LAC. It means, ultimately, that making the most disadvantaged students invisible is hiding the student populations from the poorest and most vulnerable families in our societies from the sight of policymakers and the public.

3.4.3 Criterion 3: Test Results Allow Accurate and Unbiased Monitoring of the Achievement of Curricular Learning Over Time

Any assessment program must meet minimum technical requirements to ensure that its results can be interpreted in terms of curricular learning. These technical requirements relate to test reliability and measurement error, measurement bias, and year-to-year comparability of test results.

3.4.3.1 Reliability and Measurement Error

Test results should allow for precision (reliability) in monitoring the achievement of curricular learning over time. To interpret assessment results, one needs to know the accuracy of the assessment. Not only that, but it is also of crucial importance to publish the level of accuracy: assessment methods are probabilistic, and assessment users need to have information about the probabilities associated with their results.

The minimum acceptable level of accuracy (reliability) will depend on the consequences associated with an assessment. In non- or low-stakes assessments (e.g., sample testing for monitoring purposes), it is commonly accepted that they must have internal Cronbach's alpha reliability levels $= 0.70$. In tests with higher stakes (e.g., census-based tests reporting at the school level), this indicator must be higher. Similarly, the classification error associated with performance levels may vary depending on the consequences.

The results report should account for the degree of uncertainty associated with them by indicating, for example, which score differences are statistically significant.

This can be done by using colors, asterisks, boldface text, or by grouping cases with similar results together (e.g., in the same cell of a table, regions with similar results to the national one can be shown). Another way of indicating the degree of uncertainty is to report the standard error associated with the scores.

3.4.3.2 Measurement Bias

For tests to measure curricular learning well, it is important that they be free of bias. This means that every student should have an equal opportunity to demonstrate what they have learned, so that their gender, geographical, or cultural context are not an obstacle to demonstrating their knowledge and skills. In more technical terms, it means that test scores do not contain systematic errors or interactions with specific group of students.

The bias may manifest itself as a general characteristic of the test or at the level of specific item. Overall, it may be that the average score of a test better predicts one criterion (e.g., secondary school grades) for one group of students (e.g., males) than another (e.g., females). At the item level, differential item functioning analysis (DIF analysis) is a commonly used technique to identify items that may have some bias. This analysis is complemented by a qualitative judgment on each item, carried out by professionals.

Gender bias can occur in items that are contextualized in issues that are more familiar to males than females, for example, in items about a soccer game. If a question on a test requires familiarity with soccer rules and if there are differences, for example, in knowledge between boys and girls on this subject, it will be easier for the former to answer correctly. However, this does not allow us to infer that boys attained more of the goals of the curriculum, or that they necessarily have better reading abilities. The only inference—irrelevant to curriculum policy—is that boys know more about soccer than girls. National assessments in LAC that do not yet include this type of bias analysis should consider including them.

3.4.3.3 Comparability

Countries need to monitor trends in learning over time. They must report whether learning outcomes have improved, got worse, or have been stable between assessments. To do this, it is essential that assessments be comparable, therefore, they must meet a number of technical requirements: they must measure the same assessment domain (e.g., reading), in the same student population and with equivalent samples (e.g., fourth-grade national sample), and test scores must be on the same scale. It is on this last point where validity evidence is scarce in LAC, and, therefore, comparisons of some results over time tend to be dubious.

Measuring changes in learning requires that the tests be on the same score scale. This would be simple to achieve if the exact same tests could be administered in

different years.[3] However, this is neither possible nor desirable for two main reasons: (1) the tests must be modified to be aligned with curricular updates,[4] and (2) the test items must be renewed to replace items released for publication to show what is being measured and how.

How do you measure change with constantly changing tests? The key is to make instruments that measure the same assessment domain (e.g., reading), with certain variations to accommodate curriculum updates. To this end, one part of the test can be administered in exactly the same way in different years (e.g., by repeating half of the items in both tests). The other part of the test is new (developing new items for the other half of the test). This design allows one to include both old and new items into the tests on the same score scale. This procedure, known as *equating,* applies the psychometric model of Item Response Theory (IRT). An important next step for LAC countries will be to use these types of methodologies to measure changes in learning.

There are limits to the validity of yearly comparisons between tests. When there are continuous curricular reforms that affect the fundamental elements of the assessed curriculum, these changes unavoidably influence test specifications, and the comparability of assessments administered in different years becomes more questionable. Unfortunately, there are countries in LAC that seem to be dedicated to constant cycles of curricular reforms without regard to the effect these have on the validity of inter-annual comparisons (Fig. 3.1).

3.5 Dimension of Curricular Validity of Performance Levels

To ensure assessment validity, LAC countries need to collect evidence on performance levels as measured by these assessments. Table 3.1, the Dimension of Curricular Validity of the Performance Levels, presents four priority criteria for LAC, along with the relevant validity evidence for each. These criteria are: (4) Performance levels are aligned with the curriculum, (5) Performance levels are operationalized with actual student learning in mind, (6) Performance levels describe qualitatively different stages of learning, and (7) performance levels balance stability and change within the context of dynamic curriculum policies. Table 3.2 presents a checklist with more detailed examples of evidence on this dimension.

[3] Responding to the principle: "If you want to measure change, do not change the measure."

[4] By curricular updates, we understand adjustments that do not affect the fundamental elements of the evaluated curriculum. This is, adjustments of content, skills, or competencies to be achieved in a certain grade or educational cycle. This is usually the case when making updates or curricular reforms in LAC.

The United States Federal Assessment Program **NAEP (National Assessment of Educational Progress)**

NAEP is an international benchmark for best practice in educational assessment. This is a federal assessment that is used strictly for the purpose of monitoring learning over time. Since there is no national curriculum in the United States, NAEP evaluates assessment frameworks that have been validated and agreed upon by all the states in the nation. These assessment frameworks in detail describe and justify the assessment domain of each test. Assessment domains range from the more traditional areas of reading and mathematics to more innovative areas such as economics and foreign languages. The assessment frameworks remain stable for about 10 years.

The NAEP assessment frameworks are operationalized in the test and item specifications. The specifications offer the most concrete guidelines for designing tests and writing items. These guidelines include categories traditionally used in LAC, such as content and skills. They also include categories that are more innovative, such as level of item complexity (low, medium, or high), context (e.g., theoretical or applied mathematics), item format (e.g., multiple-choice or open-ended). These documents are available online and serve as a reference for internal quality control and external audits. See mathematics example here: https://nagb.gov/naep-frameworks/mathematics.html

The alignment between tests and assessment frameworks is validated externally. Panels are formed with technical and political representation, including classroom teachers, school administrators, and other stakeholders (e.g., parents, civil society representatives), as well as curriculum and assessment specialists. This participation increases awareness and commitment to the performance levels, and contributes to their social and face validity.

NAEP uses a wide range of state-of-the-art psychometric analyses. The IRT model is used to put tests with different items on the same score scale, which allows for the interpretation of score differences as real differences in student learning. The level of uncertainty associated with the scores (standard error) is known and reported both in reports intended for the public and in technical documentation. Bias analyses are performed as part of basic testing procedures.

Fig. 3.1 International example of best practices in alignment. *Note* Prepared by the authors

3.5.1 Criteria: 4. Performance Levels Are Aligned with the Curriculum and 5. Performance Levels Are Operationalized with Actual Student Learning in Mind

There is a growing trend in LAC to report results by performance levels (also called performance standards, achievement levels, learning levels, or equivalent terms). For example, countries may report assessment results by citing percentages of students who have reached the advanced, intermediate, or basic level. Performance levels give pedagogical meaning to the results, indicating what students at different performance levels know and can do. This is essential if teachers and other educational stakeholders are to understand the results of the assessments, value them, and be able to use them for improvement.

Performance levels serve a dual purpose. On the one hand, they convey an expectation of what students are supposed to know and be able to do according to the official curriculum (criterion-referenced element, with an absolute criterion). Usually, such descriptions correspond to the curricular expectations of the grade or assessment cycle. On the other hand, performance levels are meant to describe reality, or what students actually know and can do. They show the entire distribution of students' skills and indicate their relative position on the score scale (normative element).

Performance levels should be aligned with both the curriculum and actual student learning. The descriptions associated with each level of performance should refer to the competencies, content, or skills specified in the curriculum. The highest level should correspond to the curricular expectations (key learning, terminal, or minimum learning requirements for the grade or cycle being assessed), and the lowest level should reflect what less advanced students can and do. This requires tests that include items of different levels of difficulty, from the easiest to the most difficult, so that they can discriminate well not only at the top but also at the bottom of the distribution of skills.

In LAC, it is common to observe that performance levels are set too high in relation to the actual learning achieved by students. In some countries, these results show that up to half of the students do not reach the first cut score associated with the levels. Results of this type are of limited value in informing educational policy and practice, and do little to help students who need the most support. The fact that performance levels are too demanding can be explained by a combination of factors:

(a) Performance levels rely on the curriculum alone,[5] without considering the actual learning of the students.
(b) There is a distorted view of what students actually know and can do.
(c) There is political pressure to align national performance levels with those of international assessments to make the national education look more rigorous.
(d) There is apprehension that once a lower level of performance is defined, it would be interpreted as a sufficient minimum level.
(e) The cut scores associated with performance levels are set in relative terms, based on the percentiles of the skill distribution (e.g., 50th, 75th, and 90th percentiles), without covering the lower percentiles.

[5] In LAC countries, curricula are frequently written without reference to the evidence showing what is actually taught and learned in classrooms. Consequently, these curricula often have learning objectives that could hardly be attained by large percentages of the student population. This presents an additional challenge to design assessments with performance levels that provide useful information about students who do not meet curricular expectations. It is also common that in the design of the curriculum, there is no collaboration among experts in educational measurement, and curricular experts. In such cases, the curriculum is not designed to be measurable. Therefore, it is often difficult to operationalize the curriculum with acceptable levels of validity for assessment purposes.

3.5.2 Criterion 6: Performance Levels Describe Qualitatively Different Stages of Learning

Performance levels must account for qualitatively different stages of student learning. Any rigorous classification requires clearly defined exclusive categories, and this also applies to performance levels. The competencies or skills described in one level of performance must be clearly distinct from the skills described in the next; and they must be written in simple, non-technical language, understandable to the widest possible audience. That is, the same skills should not be paraphrased differently at different levels. To achieve this, it is necessary to define a relatively small number of levels. There also needs to be sufficient distance between them on the scale of scores.

Adding more performance levels has important implications for test design, interpretation, and use of results. The more performance levels, the greater the demand for items to measure each level with the necessary accuracy, and the greater the risk of classification errors. In countries with accountability policies, classification errors have more serious repercussions than in countries that do not have such policies. Thus, if assessment results are used for accountability purposes, countries need to make sure the level of accuracy in such assessment is sufficiently high. That is, educational authorities need to be absolutely certain that a school is correctly classified as "unsatisfactory" or "underperforming," if this would affect its funding and reputation.

It is important to distinguish students who reach the lowest cut score associated with performance levels from those who do not. Based on the test results, it can be inferred that the students who do achieve the lowest performance level have acquired the competencies associated with the first cut score. In contrast, it can be inferred that students who do not reach the lowest cut score do not have such skills.

In LAC, the distinction is not always made between students who reach the lowest cut score and those who do not. Both groups are reported to be at a single performance level, as if they all had the skills associated with the first cut score. To differentiate between the two groups, it is important to introduce a default performance level that is applied to students who do not reach the first cut score and therefore do not have the skills associated with it.

3.5.3 Criterion 7: Performance Levels Balance Stability and Change in the Context of a Dynamic Curriculum Policy

Performance levels must balance stability and change to monitor curricular learning over time. Stability is necessary to make learning comparisons using the same benchmark. Change is necessary to conform to the curricular updates. In other words, performance levels need to resolve the tension of (a) being invariant to ensure that

United States Federal Assessment Program **NAEP (National Assessment of Educational Progress)**

- Reports results in four performance, or "Achievement Levels": NAEP Advanced, NAEP Proficient, NAEP Basic, and Below Basic. To do this, it sets three cut scores. Students at the "Below Basic" level are those who did not reach the lower cut point on the score scale.
- The cut scores associated with performance levels are set so that they have a gap of approximately one standard deviation on the score scale. With a gap this large, descriptions of what students know and can do are substantively different. It (what?) also ensures lower classification errors and greater reliability of the results.
- All methods and procedures for reaching final performance levels are published in the public domain and are rigorously documented. This documentation serves as the basis for regular external audits.
- To resolve the tension between stability and change in performance levels, NAEP uses a strategy of making assessments that measure two types of performance levels:
    - Main NAEP measures performance levels that are updated according to the latest curriculum changes. It is used to report changes in learning in the short term (e.g., from one evaluation to the next).
    - NAEP Long-Term Trends Assessment measures invariant performance levels that remain intact over time. It is used to measure changes in learning in the long term (e.g., over a 20-year period).

Assessment in the province of Ontario, Canada ***(Ontario Provincial Assessment Program)***

- Reports results at four performance levels, numbered 1 through 4. The additional fifth level applies to students who do not reach Level 1. This is a level that is defined by default and is described in the following way: "The student does not show enough evidence of the mastery of curriculum expectations to be categorized to Level 1".
- Level 1 describes what students with weaker learning performance know and are able to do. It corresponds to the minimum level of observed skills or learning achieved. The cut scores have been set in such a way that only around 2% of students do not reach the cut score separating Level 1 from Level 2.

Fig. 3.2 International examples of best practices for reporting results by performance levels. *Note* Prepared by the authors

results are comparable over time, and (b) being aligned with a curriculum that is regularly updated. In the context of a dynamic curriculum, more and more countries in the region will need to resolve this tension. Figure 3.2 shows an example of how to do this.

3.6 Dimension of Consequential Validity of the Assessments

The current state of development of assessments in LAC brings to the forefront the following criteria for collecting and assessing evidence of the consequential validity or impact of assessments: (8) assessment results are effectively communicated; (9)

there are formal mechanisms to support the use of assessments to improve learning; and (10) there are formal mechanisms to monitor the consequences of assessments for the education system. Table 3.1, Dimension of Consequential Validity or Impact of the Assessments, explains the validity criteria and sources of evidence. Table 3.2 presents a checklist with examples of impact validation requirements.

3.6.1 Criteria: 8. Assessment Results Are Effectively Communicated, and 9. There Are Formal Mechanisms to Support the Use of Assessments to Improve Learning

Countries implement national learning assessments in order to monitor learning and encourage improvement. The theory of action is that assessment results together with any associated information will be effectively communicated to different stakeholders (e.g., parents, teachers, principals, politicians), who will use them systematically to make better decisions. Such evidence-based decisions will have positive impact on educational policy and practice. That is, they will contribute to improving classroom teaching practices and student learning.

However, in LAC, it is common to hear that assessments have not produced the expected impact. Critics point out that a lot of data are produced, but are not useful information for decision-making; that there is a lack of an assessment culture that allows for the systematic use of such assessments; and that teachers do not use assessment information to make pedagogical decisions. The most critical, argue that the assessments are affecting education negatively by narrowing the notion of educational quality, stigmatizing schools, encouraging competition instead of cooperation between schools, making students drop out, narrowing the curriculum, stressing teachers, and so on (Falabella, 2014; Ministry of Education [MINEDUC], 2015). These effects would be more pronounced in the context of accountability policies, for example, when incentives are published and associated with school performance. Responding to these criticisms requires evidence of impact.

Communication of results is a major source of evidence to validate the impact of assessments. For these to have a positive impact, their results must be effectively communicated. That is, educational stakeholders (e.g., parents, teachers, ministry officials) need to have access to such information, understand it and value it positively, and use it effectively. There are very few studies that investigate the communication strategy of assessments in LAC (see Taut et al. (2009), and Sempé and Andrade (2017)).

The existence of formal mechanisms for using assessment results is another source of validity evidence. Examples of such mechanisms include policy guidelines that promote the use of assessment results to inform school improvement plans, monitor overall school performance, or provide feedback to teacher training programs. The effective use of assessments requires these formal mechanisms, usually absent in

LAC. Assessment programs should be responsible for promoting the use of the information they generate to improve the quality of education.

3.6.2 Criterion 10: There Are Formal Mechanisms to Monitor the Impact of Assessments on the Education System

It is key to systematically and regularly collect evidence on the consequences or impact of assessments on the education system. Such evidence should cover both positive and negative, expected and unexpected consequences. It will serve to either confirm the theory of action that guides assessments, or to modify it. The evidence will also be used to respond to the main criticisms against assessments. As long as there is evidence of impact, it will be possible to justify (or modify) the use of assessments for decision-making; such evidence is essential for providing the system with credibility.

At the global level, the evidence on the consequences of assessments is mixed. On the more positive side, the report by Mourshed et al. (2010) concludes that education systems that improve the most over time implement rigorous learning assessment policies. These assessments allow educational systems to systematically monitor learning and use results as feedback. These policies are especially relevant for countries where the quality of education is relatively low. On the other hand, there is abundant evidence of the negative impact that high stakes national assessments can have on the education system, including curricular narrowing, student and teacher stress, and even increased marginalization of vulnerable populations (Kearns, 2011; Knoester & Wayne, 2017; Segool et al., 2013).

The global evidence regarding the impact of accountability policies is also varied. These policies cover everything from the publication of school results in ranked order (league tables) to the distribution of monetary incentives to schools. In some developing countries, these policies have had a positive impact on student learning, contributed to lowering student drop-out rates, and allowed better control and supervision for education stakeholders (including parents), thus reducing corruption. However, the evidence also suggests that in other developing countries, the policies of accountability have not had the expected impact (Bruns et al., 2011).

In LAC, there are few formal mechanisms in place to collect evidence on the impact of assessments on the school system. Rather, there are isolated studies on this subject; studies that are usually financed through external funds, independent of the assessment programs.

Evidence on the impact of assessments on the education system in LAC countries is scarce but growing. The most optimistic report states that simply communicating assessment results has positive effects on student learning (de Hoyos et al., 2017). However, the evidence is usually indicative of much more moderate impact, if any. In Chile, the accountability policies have not shown the expected impact either on

improving educational policies and practices in schools (Elacqua et al., 2015), or on parents' decisions (Mizala & Urquiola, 2007) (Fig. 3.3).

3.7 Conclusions

Assessments in LAC have matured enough and have gained enough influence on the education systems and the society at large to warrant a call for validity evidence. This type of evidence is produced and used to prove that assessments effectively measure the curriculum, their results can be interpreted as performance levels, and assessments have a positive impact on the education system in general, and on student learning in particular. This evidence is necessary to give assessments political credibility and viability, and to improve their technical characteristics. It is also key for avoiding the negative costs associated with the misuse of assessment results.

This chapter presents three priority dimensions or sources of validity evidence for LAC: (1) evidence regarding the alignment of tests with the official curriculum, (2) evidence regarding the curricular validity of performance levels used to report the assessment results, and (3) evidence of consequential validity, or impact of assessments on the improvement of the education system in general, and of learning in particular. For each of these dimensions, it provides criteria and examples of evidence needed to validate the assessments.

The evidence of validation for these three dimensions in LAC is sporadic. It is in the dimension of alignment with the curriculum where the greatest amount of evidence of validation is found. There are two main challenges in this dimension. The first is that LAC countries express considerable reluctance to develop a measurable curriculum. Therefore, curricular objectives are usually formulated in very general terms, without specifying their level of difficulty or complexity. This is in contrast to practice in more technically advanced countries and education systems, where there are close professional and institutional links in curriculum and assessment design that set curricular goals in measurable terms from the outset. One reason for the high quality of assessments in the province of Ontario, Canada, is that they have a curriculum designed to be feasibly measurable. The second challenge is that curricula are constantly changing, either through reforms or updates, which makes it difficult to monitor learning achievement over time.

In the dimension of curricular validity of performance levels, countries must resolve technical–political tensions in order to measure the achievement of curricular learning. How many levels should be set? How demanding should these levels be? Where should these levels be set on the scale of scores?

Another tension that countries need to solve arises from the United Nations' Sustainable Development Goals. Goal 4 of "Ensuring inclusive, equitable and quality education and promote lifelong learning opportunities for all" will be measured in part by indicator 4.1.1, "Proportion of children and young people (a) in second or third grade, (b) at the end of primary school and (c) at the end of lower secondary school achieving at least a minimum proficiency level in reading and mathematics, by

The United States Federal Assessment Program **NAEP (National Assessment of Educational Progress)**

NAEP has written communication policies that identify responsibilities for reporting and disseminating results. NAEP strategic communication plans identify principles and priorities for disseminating assessment results. These plans are developed by the NAEP's Reporting and Dissemination Committee from the National Assessment Governing Board (NAGB) and are available online. The NAEP Validity Studies Panel reviews aspects of validity and uses of NAEP publications. It also funds external studies to collect evidence on this aspect of validity. NAEP has a validity study schedule that covers evaluating the consequences of the reporting of results.

U.S. State Assessment Programs

There is a great deal of research in the United States on the impact of assessments on teaching practices and student learning. These investigations are made possible by the existence of public and private funds that prioritize funding for impact validity studies. Thus, for example, a study of the impact of state assessments found that when performance levels had more positive names (labels), students were more likely to decide to continue their education at the post-secondary level (Papay, Murnane, & Willett, 2016). In New York, a study of the impact that the new statewide fourth grade assessments had on teacher turnover found that there was fewer turnovers both in this grade after their introduction and in relation to other grades not assessed (Boyd, Lankford, Loeb, & Wyckoff, 2008). In Texas, a study of the impact of statewide assessment found that the policies of accountability had more positive impact in schools with low average levels of achievement in mathematics, where it was observed that students took more mathematics courses and achieved higher levels of performance over time. Surprisingly, the impact was reversed in schools with higher average levels of performance (Deming, Cohodes, Jennings, & Jencks, 2016).

Ontario Provincial Assessment Program

The EQAO (Education Quality and Accountability Office), an organization in charge of assessments, regularly collects evidence on how its assessments are used and what impact they have on the education system. It implements the following activities:

- Internal reviews (e.g., through forums) to gather information about tests and how they are conducted in schools; the relevance of reports to educational accountability and improvement; and the impact of testing on teacher training. These forums are conducted jointly by the Assessment Advisory Committee and more than 20 organized community interest groups representing public and private school principals, supervisory agents, teachers, boards of education and trustees, parents, and students.

- Regular external audits that cover the overall assessment process, including its impact on the education system. To this end, consultations are held with stakeholders and the general public. The external audit report is made available for public comment.

As a result of internal and external reviews, the EQAO makes commitments to action to improve assessments.

Fig. 3.3 International best practices of consequential validity of assessments. *Note* Prepared by the authors

gender." When defining performance levels for their national assessments, countries should take into consideration the minimum proficiency level determined internationally by the global teams defining standards for reporting on SDG indicator 4.1.1. However, they should do this without compromising the alignment of performance levels with their own national curriculum.

The effective use of assessments requires creating an assessment culture where education stakeholders can access, understand, value, and use assessments for improvement. Forming this assessment culture is a pending challenge in LAC. The centrality of this point contrasts with the marginal budget that many countries allocate to the communication of results and general information on assessments. It also contrasts with the lack of opportunities (e.g., workshops, courses) for teachers, managers, or ministry officials to reflect on the obtained results and possible ways to improve them in the future.

Another important step is to move forward in consolidating an agenda for validating the consequences of the assessments. Collecting this type of evidence is not part of the countries' working agenda. This puts the credibility of the assessments at risk and has important associated costs, such as the costs of decisions based on erroneous assumptions. One example would be assuming that teachers will use results published in the reports to improve their teaching practices, when in fact the reports do not reach their audience and, when they do, teachers do not understand them. Alternatively, an assumption that teachers will improve their teaching practices when the results show that most of their students are clustered at the lowest performance level is a weak one. Without information on the students who are showing least progress, it is implausible that there will be improvements in the pedagogical practices of teachers and in the learning of these students.

A challenge for Latin American countries is to resolve the tensions inherent in meeting different validation criteria. For example, there exists a conflict between evaluating curricular objectives and assessing what all students know and can do, including those who lag behind; or between reporting invariant performance levels and adjusting the levels according to curricular updates.

It is important for countries to prioritize validity dimensions and criteria, depending on the degree of maturity of their assessments and the local context. For the countries that are currently introducing assessments to their education systems, the dimension of assessment alignment with the curriculum would be a priority with the minimal goal of producing technical documentation on test specifications. In countries that are starting to report results by performance levels, the priority should be for these reports to describe what students know and can do with regards to curricular objectives and actual student learning. Countries that have already put their assessments in place should focus on evidence of validity regarding the effective communication and impact of the assessments.

The cost of validating the assessments is lower than the cost of not validating them at all. Expenses associated with conducting an external audit, for example, may seem considerable compared to the annual budget of an assessment program. However, this cost is unlikely to exceed 5% of the annual budget, and its benefits may go far beyond one assessment cycle. The cost of not having evidence of validity, however,

can be much higher; for example, negative misinterpretation of the results by the press can lead to a political crisis and even force educational officials to resign.

The dimensions and validity criteria presented in this chapter can be used for different purposes:

- *To carry out an internal review of the methods and procedures used.* Assessment program teams can use them to define technical standards to be met and to define their internal work routines for test validation.
- *To define a validity study agenda.* For example, it could be decided that in a current year, a validity study will address the alignment of the tests with the curriculum, and that in the following year it will focus on the impact of the assessments on teaching practices.
- *To provide technical assistance in assessment to countries.* There is strong interest from the international community to strengthen learning assessment programs in the countries of the region. The validation criteria presented here may be useful in identifying weaknesses and technical assistance needs (for example, to review the evidence regarding effective communication of assessments).
- *To guide external audits.* Assessment programs should not only receive internal evaluations of their procedures and results; they must also go through external review. Claiming that a student, school, or education system has reached a certain level of performance is a statement that should be auditable. That is, it must be independently verifiable, in order to justify the decisions made on the basis of these results. Conducting external audits is a regular practice at NAEP (the USA) and the Provincial Assessment Program (Ontario, Canada). In LAC, there are several notable but still scarce examples of this type of practice (see the chapter by Valverde and Ramírez in the present volume).

In LAC, validation reviews, where they exist, tend to be internal and more process-focused—that is, the assessment programs themselves collect or judge evidence of validity. Being more focused on processes, they are not conducive to creating a culture of external audits of their products. It is assumed that if the process was carried out as planned (implementation fidelity), the product will be adequate as well. Countries would benefit from establishing a regular external audit policy that is transparent to citizens, including better documentation of the processes used and assessment instruments produced. These audits should be seen as an opportunity to improve the assessments.

The degree of maturity of assessments in LAC and of the level of influence they have gained in the public sphere requires evidence of validity: evidence to support the claim that the assessments are fulfilling their purpose of monitoring the achievement of curricular objectives and encouraging improvement. The dimensions and criteria proposed here seek to support countries in their systematic search for such evidence. Following these criteria is expected to give greater credibility, as well as political and technical feasibility to the assessments, in the interest of a better education for all.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2008). The impact of assessment and accountability on teacher recruitment and retention: Are there unintended consequences? *Public Finance Review, 36*(1), 88–111. https://doi.org/10.1177/1091142106293446

Bruns, B., Filmer, D., & Patrinos, H. A. (2011). *Making schools work: New evidence on accountability reforms.* World Bank. https://doi.org/10.1596/978-0-8213-8679-8

Darling-Hammond, L., Herman, J., Pellegrino, J., Abedi, J., Aber, J. L., Baker, E., & Steele, C. M. (2013). *Criteria for high-quality assessment.* Stanford Center for Opportunity Policy in Education.

de Hoyos, R., García-Moreno, V., & Patrinos, H. A. (2017). The impact of an accountability intervention with diagnostic feedback: evidence from Mexico. *Economics of Education Review, 58*(C), 123–140. https://doi.org/10.1016/j.econedurev.2017.03.007

Deming, D. J., Cohodes, S., Jennings, J., & Jencks, C. (2016). When does accountability work? The Texas system had mixed effects on college graduation rates and future earnings. *Education next, 16*(1), 71–76.

Elacqua, G., Martínez, M., Santos, H., & Urbina, D. (2015). Short-run effects of accountability pressures on teacher policies and practices in the Chilean voucher system in Santiago, Chile. *School Effectiveness and School Improvement, 27*(3), 385–405. https://doi.org/10.1080/09243453.2015.1086383

Falabella, A. (2014). The performing school: The effects of market and accountability policies. *Education Policy Analysis Archives, 22*(70), 1–29. https://doi.org/10.14507/epaa.v22n70.2014

Kearns, L. L. (2011). High-stakes standardized testing and marginalized youth: An examination of the impact on those who fail. *Canadian Journal of Education, 34*(2), 112–130.

Knoester, M., & Wayne, A. (2017). Standardized testing and school segregation: Like tinder for fire? *Race, Ethnicity & Education, 20*(1), 1–14. https://doi.org/10.1080/13613324.2015.1121474

Ministry of Education. (2015). *Towards a complete and balanced system of learning assessment in Chile. Report task force for the revision of the Simce.* Retrieved from http://www.mineduc.cl/wp-content/uploads/sites/19/2015/11/Informe-Equipo-de-Tarea-Revisi%C3%B3n-Simce.pdf

Mitroff, I. I., & Featheringham, T. (1974). On systemic problem solving and the error of the third kind. *Behavioral Science, 19*(6), 383–393.

Mizala, A., & Urquiola, M. (2007). School markets: the impact of information approximating schools' effectiveness. *Journal of Development Economics, 103*(C), 313–335. https://doi.org/10.3386/w13676

Mourshed, M., Chijioke, C., & Barber, M. (2010). *How the world's most improved school systems keep getting better.* McKinsey & Company.

Papay, J. P., Murnane, R. J., & Willett, J. B. (2016). The impact of test score labels on human-capital investment decisions. *Journal of Human Resources, 51*(2), 357–388.

Segool, N. K., Carlson, J. S., Goforth, A. N., Von der Embse, N., & Barterian, J. A. (2013). Heightened test anxiety among young children: Elementary school students' anxious responses to high-stakes testing. *Psychology in the Schools, 50*(5), 489–499. https://doi.org/10.1002/pits.21689

Sempé, L., & Andrade. P. (2017). *Final report: Evaluation of the use of census reports of students in the school.* GRADE/FORGE.

Taut, S., Cortés, F., Sebastian, C., & Preiss, D. (2009). Evaluating school and parent reports of the national student achievement testing system (Simce) in Chile: Access, comprehension and use. *Evaluation & Program Planning, 32*(2), 129–137. https://doi.org/10.1016/j.evalprogplan.2008.10.004

UNESCO Institute for Statistics, & Australian Council for Educational Research. (2017). *Principles of good practice in learning assessment.* Recovered from http://uis.unesco.org/sites/default/files/documents/principles-good-practice-learning-assessments-2017-en.pdf

Valverde, G., & Ramírez, M. J. (2019). Contemporary practices in the curricular validation of national learning assessments in Latin America: A comparative study of cases from Chile, Mexico and Peru. In J. Manzi, M. R. García, & S. Taut (2019), *Validity of educational assessment in Chile and Latin America.* Ediciones UC.

María José Ramírez Psychologist from the Pontificia Universidad Católica de Chile and Doctor in Education from Boston College, USA. Currently she is an international consultant in education. Her areas of interest include learning assessment and quality of education. Contact: mariajose.ramirez001@gmail.com

Gilbert A. Valverde He has a degree in Philosophy from the University of Costa Rica and a Ph.D. in Comparative and International Education from The University of Chicago, USA. He is Dean of International Education and Vice-Provost for Global Strategy at the University at Albany, State University of New York and a member of the faculty in the Department of Educational Policy and Leadership. His areas of interest include the international comparative study of assessment policy and curriculum, educational measurement, and large-scale international testing. Contact: gvalverde@albany.edu