Chapter 2
Is Validation a Luxury or an Indispensable Asset for Educational Assessment Systems?

Sandy Taut and Siugmin Lay

2.1 The Modern Definition of Validity and Validation

Most Latin American education systems have developed assessment programs, and their results constitute a significant reference for discussion on education in each country. These programs range from standardized national and international student assessment systems to teacher evaluation programs and university entrance exams. All these types of evaluations are included in the book you hold in your hands. The information that these assessments produce is expected to be used to inform different types of decisions by different stakeholders. For example, a standardized student assessment system can serve, in a first case, to monitor student competencies at different levels of the education system or, in a second case, to provide diagnostic and comparative information on what students in an institution know and can do. Consequently, users of assessment information are, in the first case, national, regional, and local policymakers and legislators, and in the second case, administrators, school leaders, teachers, students, and their families. This book aims to address a fundamental question in this regard: is the information provided by these tests and evaluation programs of sufficient quality to make these proposed interpretations and uses in fact useful and justified?

This question refers to the standards of quality that a test or assessment system must meet. All contributions to this book refer to the Standards for Educational and Psychological Testing (American Educational Research Association [AERA] et al.,

S. Taut (✉)
Bavarian State Office for Schools, Gunzenhausen, Germany
e-mail: staut@uc.cl

S. Lay
Pontificia Universidad Católica de Chile, Santiago, Chile
e-mail: splay@uc.cl

J. Manzi et al. (eds.), *Validity of Educational Assessments in Chile and Latin America*,
https://doi.org/10.1007/978-3-030-78390-7_2

2014,[1] hereinafter referred to as "the Standards"), which describe such expectations for educational (and psychological) testing and assessment systems.

The most central of standards—in fact, the first one—points to the concept of validity. According to the Standards, *validity refers to how much evidence exists to support particular interpretations of test scores for specific uses of tests.* Therefore, the validity of a test or evaluation system does not exist by itself, but only in relation to specific interpretations and intended uses. It has been a long time since validity was understood as a property of the test or evaluation. The statement "this test has high validity" can no longer be said when we adhere to a modern understanding of validity and validation, one that is proposed in the Standards and in this book.

For instance, can a school leader interpret the results of an eighth grade standardized mathematics test as diagnostic information about the skills his or her students have acquired up to eighth grade? Are the test results useful for drawing conclusions about teacher quality in mathematics in a given classroom? Can the school leader use the test results to compare his or her school to other schools? These are questions that relate to the validity of a student assessment system, and for each of the interpretations and uses mentioned previously, validity must be examined systematically. It is necessary to know how much evidence exists to support these interpretations and uses, and the extent to which the evidence tells us that these interpretations and uses are not justified and should, in fact, be avoided.

The process of examining validity and gathering evidence to support (or refute) a validity argument is called validation (Kane, 2006). Validation begins by *explicitly stating the intended interpretations and uses.* As can be seen from the mathematics test example above, it is not always easy to identify or agree on what are the intended interpretations and uses of the test scores, and what would be unintended uses in a given context. An educational assessment system must often serve multiple purposes. It is also common for various stakeholders to have conflicting interests about the interpretations and uses of the obtained results. However, having clarity about the overall purposes of the tests and evaluations, the main users, and the interpretations, uses, as well as intended and unintended consequences (perhaps summarized in the form of a logic model or theory of action, see Santelices (2021) in this book) is of great importance as a basis for validation work to be productive.

In the same logic, *a detailed description of the construct to be measured is needed*, because, as a conceptual framework, it tells us what interpretations may or may not be derived from assessment scores. For example, if we want to assess student achievement in a given area, then we must conceptually define what achievement means and implies in this context, so that the assessment really focuses on important aspects of this construct. In this case, one way to conceptually define achievement is through the curricular system that establishes learning objectives and competencies by educational level. If the assessment measures content or competencies not included in the curriculum, then the assessment result could no longer be interpreted as an indicator of student achievement in that particular area.

[1] The 2018 version of the Standards corresponds to the Spanish translation of the 2014 Standards edition; thus, it does not refer to a new version of the Standards.

Therefore, the validation process should *explain the propositions that must be met in order for test scores to be interpreted as intended.* Each interpretation for a specific use must be supported by evidence. Because testing and assessment systems have dynamic contexts and undergo frequent changes and/or adaptations, validation should be an ongoing process in which evidence is regularly collected in order to re-examine and strengthen a line of argument about the validity of each interpretation or use.

Different sources of evidence are necessary and appropriate to generate and contrast a validity argument. According to the Standards, these include: (1) evidence based on test content, (2) evidence based on response processes, (3) evidence based on internal structure, (4) evidence based on relations to other variables, and (5) evidence based on test consequences. Each one of these sources of evidence will be explained in greater detail later in this chapter. Although each of these sources of evidence contributes separately to examining the propositions that underlie certain interpretations of the test, these do not constitute different types of validity, since validity is now understood as a *unitary concept.*

The final, and the most challenging step in validation is to integrate the different sources of evidence into a *coherent argument about the validity of a given interpretation and use* of the test or assessment results. For example, one might want to validate the use of scores in a math test to compare the math skills of students from different schools in an education system. For this use, we could collect evidence that the test content is reflecting the curriculum implemented in the schools; examine that the test's internal structure shows the different sub-domains of mathematical competence; ensure—through studies of response processes—that students understand the test items in the intended way; examine relations to other relevant variables (e.g., grades in mathematics, at school level); and implement studies on the intended and unintended consequences of testing on students, teachers, and schools. Then, we would have to integrate the evidence and reach a conclusion about the possibility of interpreting these math test scores as a basis for making comparisons between schools.

While this book's focus is on validity as the central aspect of ensuring a useful and effective testing and assessment system that fulfills its purposes, the Standards indicate that reliability/precision and fairness are additional, important aspects of quality. On the one hand, reliability/precision refers to the consistency of the score between different applications, forms, or parts of the test, and is typically reported as a reliability index and a standard error of measurement. Reliability/precision is a necessary, but not sufficient, condition for validity. That is, even when test scores are highly consistent, this does not imply that the instrument is able to diagnose, predict, or promote appropriate decisions.

Fairness, on the other hand, points to the ethical obligation to give all examinees or test takers equal access to demonstrate their ability in the test without obstacles. Fairness also impacts on validity and reliability/precision, particularly when the test is biased towards certain subgroups, i.e., when—despite having the same level of the attribute to be measured—a specific subgroup obtains a lower score in the assessment than the rest of those being tested.

We will now briefly describe the history of the concept of validity in order to explain how it became a unitary concept, how it evolved into an approach about interpretations and intended uses, and how it became an argument-based approach.

2.2 The History of Validity

2.2.1 Early Definitions of Validity

Until the early 1920s, there were no major questions regarding the quality of the measurements used in research and their relationship to the constructs to be measured. This issue was only addressed in the course of the measurement industry boom for selection purposes in World War I. In order to achieve a consensus on the procedures and terms used in measurement, back then validity was defined as the degree to which a test or assessment measures what it claims to measure (Ruch, 1924). This corresponds to the classical definition of validity, and at that time referred to the measurement itself and not to the interpretation of the results or the use of the test.

Based on this definition, two different approaches were developed in order to establish proof of validity about a test. On the one hand, a logical analysis of test content by experts could indicate whether the evaluation was measuring what it claimed to measure. On the other hand, the correlation between the test and what it was intended to measure was a method that could provide empirical evidence of validity. The development of these different approaches to establish test validity prompted the definition of validity to evolve, and various types or aspects of it began to be differentiated. For instance, Cronbach (1949) differentiated between logical and empirical validity, while the APA in its first Standards (American Psychological Association, 1952; American Psychological Association et al., 1954) went further and distinguished four types of validity: predictive, concurrent, content, and construct validity. These four types of validity also implied different types of analyses to provide relevant evidence.

Predictive validity was to be established by correlating test results with a variable allegedly predicted by the measured construct, assessed some time later. Concurrent validity was to be established by evaluating the relationship between the results of a test and a variable measuring the same construct, at the same point in time. Content validity was to be established when the test was considered a sample of the universe that operationally defines the variable to be measured (e.g., evaluating writing speed through a writing exercise). Finally, construct validity was to be established when performance on the test was presumed to rest upon a hypothetical attribute or quality. That is, unlike content validity, this attribute would not be defined operationally but theoretically. Therefore, in construct validity, the interpretation of what is measured becomes relevant. In this case, no particular type of analysis existed by default, but theory-driven predictions had to be generated and tested, since neither logical analysis nor empirical evidence was sufficient.

Later, in the second and third editions of the Standards (APA et al., 1966, 1974), content and construct validity were preserved, while predictive and concurrent validity were brought together under a single category: criterion-related validity.

In sum, validity initially did not refer to the interpretation of results or the use of the test, but to the measurement instrument itself. Additionally, four types of validity were differentiated: predictive, concurrent, content, and construct validity (or three, when considering predictive and concurrent validity as part of criterion-related validity). However, these different aspects or approaches to validity were paving the path for alternative conceptions of validity.

2.2.2 Messick's Unified Construct of Validity

Messick (1989, 1994) criticized this traditional view on validity, considering it incomplete and fragmented, while making considerable contributions to the theory of validity. First, Messick indicates that validity no longer refers to a property of the test regarding whether or not it measures what it claims to measure, but validation is a scientific process that must focus on the meaning or interpretation of test scores. This change of focus is based on the fact that test scores do not exclusively depend on the questions and items in the test, but also on the examinees and the context. Thus, what needs to be valid is the scores' meaning or interpretation, as well as any practical implications that arise from the test (Messick, 1995).

Moreover, Messick recognizes that the study of validity not only plays a scientific role in measurement, but it has a social value and plays a political role too, by impacting on judgments and decisions outside the measurement realm. Because of this, validity must take into consideration not only the scores' meaning as a basis for action, but also the social consequences of the use of the test. In other words, Messick integrates an ethical component into validity.

Another contribution of Messick is to conceive validity as a unitary concept of construct validity. The author identifies two main threats to the valid interpretation of test scores: factors irrelevant to the construct and under-representation of the construct. Through the unitary concept of construct validity, Messick integrates logical analysis and empirical evidence into the validation process. Therefore, for this author, validity is no longer the result of a single analytical or empirical study, and he calls for as much analysis and evidence as possible to be gathered before establishing a validity claim. As can be seen, he no longer refers to types of validity, but to types of evidence that support validity.

Therefore, according to Messick, the types of evidence must be combined, and finally, based on the evidence gathered, a judgment must be made as to whether or not the test scores can be interpreted as intended and whether the test can be used for the purposes for which it was intended. Messick's unified concept of validity considers the aspects of criterion, content, and consequences within a construct validity framework. Therefore, construct validity is the validity of the test that looks into the meaning of the score and the consequences of test use. Messick's contributions to validity have been fundamental, and they strongly set the foundations for the fifth edition of the Standards (AERA et al., 1999) and inspired Kane's more recent contributions (2015, 2016).

2.2.3 Cronbach's Notion of Validation

Cronbach, like Messick, emphasized the importance of constructs in validation. Together with Meehl, he pointed out that a construct is defined by a nomological network that would relate this construct to other constructs as well as to observable measures (Cronbach & Meehl, 1955). These expected relationships make construct validation possible. If in a validation study, those expected interrelationships are not found, then there would be a flaw in the posited interpretation of the test score or in the nomological network. In the first case, construct validity would not be supported by the evidence; in the second, the nomological network would have to be modified, and therefore, the construct would change. In this sense, the theory behind the construct is central, since it establishes the relationships between the different constructs, guiding the hypotheses that will be tested in the construct validation studies.

In the same vein, Cronbach (1989) advocated a strong validation program. He described a weak program as one in which the researcher does not define a construct to lead the research, but rather the investigation aims to obtain some result without having a clear course. Whereas a strong validation program has the construct at its core, from which construct hypotheses are derived, and different types of relevant evidence are gathered to test them. In other words, establishing a construct in the strong validation program allows a more focused validation, while in the weak program this focus is lost. According to Cronbach, a strong program with well-defined constructs must be at the core of validation.

Finally, Cronbach (1989) also indicates that there is an obligation to review the appropriateness of a test's consequences and prevent them from being adverse. However, unlike Messick, Cronbach does not make any statements as to whether or not the consequences should be included in the validation process itself.

2.2.4 Kane's Contribution on Validity Arguments

Kane complements Cronbach's and Messick's contributions. While Messick emphasized the importance of making a general judgment of validity based on different types of evidence, Kane developed in more detail how such a judgment can be constructed and tested.

Kane (1992) stated that the interpretation of the scores and the use of the test always involve reasoning based on interpretive and use arguments. This reasoning describes the propositions that lead from the test score—the premise—to the interpretations and decisions based on that score—the conclusions. Hence, for Kane, the validation process consists of two steps: (1) all propositions at the basis of the test scores' interpretations and uses must be converted into an explicit argument for interpretation and use, and (2) the plausibility of these propositions must be critically assessed in a validity argument. This way, propositions will lead the search for adequate methods and evidence that will allow to support or reject any given interpretation and use of the test (Kane, 2015, 2016).

In summary, Kane emphasizes the importance of propositions at the base of interpreting and using test scores, and elaborates the validation process as an argumentation exercise. Kane's contribution has been incorporated into the most recent editions of the Standards (AERA et al., 1999, 2014).

2.2.5 Are Consequences of Testing Part of Test Validation?

Several scholars have highlighted the importance of evaluating measurement consequences in validation processes, in particular, those unintended consequences that are generated as a result of the assessment (Kane, 2013; Lane & Stone, 2002; Lane et al., 1998; Linn, 1997; Messick, 1989, 1995; Shepard, 1997). This issue is especially relevant in high-stakes contexts, since research has shown that several undesirable processes often occur in these systems that distort the results, for instance, the excessive preparation on the test content that can lead to a narrowing of the curriculum (Brennan, 2006; Kulik et al., 1984; Shepard, 2002). As Kane (2006) points out, high-stakes assessment programs must be considered as educational interventions that require thorough evaluation, like any other educational program or major reform attempt (see also Zumbo, 2009).

Lane and her colleagues (1998) argue that it is essential to include both intended and unintended consequences of educational assessments as part of validation studies. The authors argue that these studies should consider the perspectives of all possible stakeholders in the assessment, from the community and policy makers, to students and their families (Linn, 1997). Moreover, this evaluation of the effects should also be conducted at various levels of analysis: at a program level, at a school district level, and at school and classroom levels (Lane & Stone, 2002).

However, there is controversy on this issue, as other researchers have argued that while the consequences of assessments are important to consider, they should not be part of the validity concept (Borsboom et al., 2004; Cizek, 2012; Maguire et al., 1994; Popham, 1997; Wiley, 1991). For instance, Popham (1997) states that validity must take into account the accuracy of inferences that are made based on test scores, i.e., validity refers to the meaning of the scores. The consequences would be a step beyond inferences. Therefore, while examining the consequences is relevant, it should not be part of the validation process. Furthermore, Cizek (2012) states that scientific and ethical arguments are logically incompatible, and therefore, validity must refer to the inferences of the test scores, while the use—and therefore the consequences—would only be part of the justification for the use of the test. In other words, validity would not depend on the use of the test. In this sense, these authors choose to leave the analysis of a test's consequences out of the validation process.

However, the Standards clearly position themselves in favor of including the consequences in the validation. In the following section, the Standards' recommendations are described in more detail since they represent a key reference in the educational measurement context.

2.3 The Standards for Educational and Psychological Testing

According to the most recent edition of the Standards, validity refers to "the degree to which evidence and theory support interpretations of test scores for proposed uses of tests" (AERA et al., 2014, p. 11). The first standard of validity expresses in the following way what is—in general—expected as the basis for any validation exercise: "Clear articulation of each intended test score interpretation for a specified use should be set forth, and appropriate validity evidence in support of each intended interpretation should be provided" (Standard 1.0; AERA et al., 2014, p. 23).

First, the Standards describe how the proposed interpretations and uses must be articulated. These should not only be clearly described, but also the population for which the test is intended must be defined, along with explicitly stating unforeseen interpretations, and/or interpretations that have not been sufficiently supported by evidence. If, following the interpretation and use of test scores, specific effects or benefits are expected, these should be supported by relevant evidence. This also applies to indirect benefits that are expected as a result of the assessment system in question. The degree to which the preparation for the test would change, or would have no effect, on the test result must also be explained. Finally, the samples and contexts in which validity studies are implemented must be described in sufficient detail to judge their quality and relevance.

In addition, the multiple interpretations and uses of the test must be validated independently of each other. In principle, it seems difficult for an assessment to have multiple purposes and uses, but in practice this is not unusual. Nonetheless, each of them must be supported by sufficient evidence of validity to be justified.

Then, the Standards distinguish between different *types of evidence of validity* that can be collected and that will be *more or less relevant depending on the interpretation and use that we want to give to the test results*. Each type of evidence is described below in more detail.

2.3.1 Evidence Based on Test Content

> When the rationale for test score interpretation for a given use rests in part on the appropriateness of test content, the procedures followed in specifying and generating test content should be described and justified with reference to the intended population to be tested and the construct the test is intended to measure or the domain it is intended to represent. If the definition of the content sampled incorporates criteria such as importance, frequency, or criticality, these criteria should also be clearly explained and justified (Standard 1.11; AERA et al., 2014, p. 26).

The test content consists of the topics, wording, and format of the test items, tasks, and questions. Content-based evidence refers to the consistency between the test content and the content domain or the construct being measured. This type of evidence also analyzes whether the content domain is relevant to the interpretation we want to give to the test scores. This highlights the importance of the content domain specification, since it describes and classifies the content areas and types of items to be used. A test may fail to capture some important aspect of the construct to be measured, thus the construct would be underrepresented. For example, a test that aims to measure reading comprehension should consider a variety of texts and reading materials, in order to cover reading comprehension of all possible types of texts (e.g., essays, news, etc.). At the same time, test results could be affected by sources of variance irrelevant to the construct to be measured, as would be the case if the test had very long reading texts that required the examinees to have a great capacity to memorize. In this case, we would be evaluating not only reading comprehension, but also memory.

This type of evidence may involve logical or empirical analyses. The use of expert judgment can support the search for evidence based on test content, and to assist in the identification of possible inconsistencies between test content and the content domain. In the example above, language teachers could judge how representative of the curricular standards the texts and questions are on a reading comprehension test.

The search for this type of evidence is of utmost importance in situations where a test that was designed for a given use is to be used for a new purpose, since the appropriateness of the construct to be measured is related to the inferences that are made based on test scores. Evidence based on test content is also useful for addressing score differences between subgroups of examinees, as these may account for sources of variance that are irrelevant to the content domain. For example, an employment test that uses more advanced vocabulary than required for the job position may create an unfair disadvantage for applicants whose first language is not English, accounting for a source of variance irrelevant to the construct.

2.3.2 Evidence Based on Response Processes

> If the rationale for score interpretation for a given use depends on premises about the psychological processes or cognitive operations of test takers, then theoretical or empirical evidence in support of those premises should be provided. When statements about the processes employed by observers or scorers are part of the argument for validity, similar information should be provided (Standard 1.12; AERA et al., 2014, p. 26).

Another type of evidence is that which analyzes the consistency between the construct and the response process of test takers. For instance, in the case of a reading comprehension test, it is expected that there will indeed be an understanding of the words used and a global understanding of the texts presented, and not that the nature of the examinees' responses reflects a different mental process. To collect this type of evidence, we can ask a sample of examinees about their response strategies, examine their response times or eye movements, or look into the relations between different parts of the test.

Evidence based on response processes can also be used to detect measurement bias, as it can shed light on differences in the interpretation of scores between subgroups of respondents. This type of evidence could inform about capacities that may be influencing subgroups' performance in different ways.

This evidence can be gathered not only from examinees, but also from reviewers who can provide useful and relevant information. In the event that the test is scored or corrected by an external party, we can evaluate the consistency between the scorer's correction processes and the criterion to be used according to the construct to be measured. For example, among raters who correct an essay, we must ensure that their grading is not influenced by the quality of the examinee's handwriting, but that they use the criterion that refers to the construct they must evaluate.

2.3.3 Evidence Based on Internal Structure

> If the rationale for a test score interpretation for a given use depends on premises about the relationships among test items or among parts of the test, evidence concerning the internal structure of the test should be provided (Standard 1.13; AERA et al., 2014, pp. 26–27).

Evidence based on the internal structure points to the degree of consistency in the relationship between the test items and the defined components of the construct to be measured. For instance, if the conceptual framework of the test defines the construct to be measured as unidimensional, then this must be reflected in item homogeneity. However, if the construct was multidimensional, the item response pattern should reflect this. For example, in a science test that assesses knowledge in the natural sciences, including chemistry, physics, and biology, one would expect chemistry items to be related to each other more strongly than with physics and biology items. There are different statistical techniques that allow us to examine and check if one or more factors are associated with the items, such as exploratory factor analysis and confirmatory factor analysis. In the case of the science test, we would expect that the exploratory factor analysis would account for three factors according to the sub-dimensions (chemistry, physics, and biology).

2.3.4 Evidence Based on Relations to Other Variables

> In many cases, the intended interpretation for a given use implies that the construct should be related to some other variables, and as a result, analyses of the relationship of test scores to variables external to the test provide another important source of validity evidence (AERA et al., 2014, p. 16).

Evidence based on relationships with other variables takes on special relevance when the intended interpretation for the use of a test assumes a relationship between the construct and another variable external to the test. This external variable can be a similar or a different construct to the one intended to be measured, a criterion that the test should be able to predict, another test that measures the same construct, etc. For example, one might expect that in a test of critical thinking skills, results would depend on the intensity of instruction and practice with this type of problems. In this case, we could obtain evidence based on relationships with other variables by evaluating the relationship between the test score and indicators about the quality or frequency of the examinee's instruction in this type of skill.

We can also analyze the relationship between the result of a test and other measurements of the same construct or variables that are similar to the construct to be measured. In this case, the evidence would be convergent, since the two measurements are expected to be related. For example, the results of a multiple choice test that assesses knowledge in chemistry should be related to other tests that also assess knowledge in that discipline, but in a different format, such as tests using open-ended questions. On the contrary, we could evaluate the relationship between the test score and other measurements of variables that are theoretically different and should not be associated with the construct. The evidence in this case would be discriminant, as we would expect the relationship between the two to be low or non-existent. In the case of the chemistry test, its results should be less related to the result of tests in more distant disciplines, such as mathematics or history.

The test–criterion relationship should be evaluated when the test is expected to predict an attribute operationally different from the test, as in the case of an employment test to select suitable applicants for a particular position. In this example, it may be important to gather evidence on whether performance on the test effectively predicts subsequent performance in that position. This would be a predictive study, as the test scores are expected to predict the criterion scores obtained later (i.e., subsequent performance). In cases where alternative measures are to be developed to assessments already approved for a given construct, the designed study would be concurrent, since both tests must be carried out at approximately the same time in order to avoid differences that could be explained by the time lag. Whenever the test–criterion relationship is to be studied, we should report information on the criterion's technical quality and appropriateness, since the credibility and usefulness of these studies depend on this information. In order to gather this type of evidence is central to establish an appropriate criterion and to measure it by applying high technical standards.

2.3.5 Evidence Based on Consequences of Testing

> Decisions about test use are appropriately informed by validity evidence about intended test score interpretations for a given use, by evidence evaluating additional claims about consequences of test use that do not follow directly from test score interpretations, and by value judgments about unintended positive and negative consequences of test use (AERA et al., 2014, p. 21).

Test use entails several consequences that may or may not be aligned with the interpretation and use of the test initially intended by the test developer. First of all, evidence based on consequences of the test use must analyze whether the effect or consequence of the assessment is consistent with the intended test interpretation and use. For example, if it is claimed that a math test can be used by teachers to diagnose students' needs for remediation classes regarding specific sub-areas, then evidence should be presented as to whether the test results actually provided useful diagnostic information as perceived by teachers.

However, there may be other anticipated consequences of the use of the test that go beyond the direct interpretation of the score. It could be expected that a test primarily used for admission to higher education might indirectly improve teaching quality via teachers taking responsibility for the learning achievements of their students (i.e., accountability). This consequence of the use of the test should then also be validated.

Additionally, there may be unforeseen consequences that may or may not affect the validity of the interpretation of the results. This would depend on whether or not these consequences originate from a source of error in the interpretation of the test score. In the example of the selection test, a subgroup of test takers (e.g., depending on age) may have a higher pass rate than the rest of test takers, and this would constitute an unintended consequence. If at the basis of this consequence, it is found that there are components in the test that negatively impact the score obtained by this subgroup, and that these are irrelevant to the construct to be measured, this consequence would invalidate the interpretation and use of the test (e.g., items in the test related to technology that are not part of the construct to be measured, and that favor younger test takers because they are more familiar than the rest of the test takers with this particular topic). However, if at the base of this consequence, we find in the student population an unequal distribution of the knowledge to be measured and this knowledge is relevant to the construct to be measured, then this difference would not invalidate the interpretation of the test results.

In this regard, it is essential to differentiate between (a) consequences arising from an error in the interpretation of the test score for an intended use and (b) consequences arising from another source unrelated to the test itself. The latter may influence the decision to use or not to use the test, but they do not directly invalidate the interpretation and use of the test. For instance, another unintended consequence would be that schools neglect the teaching of some subjects to focus only on those assessed in the higher education entrance test. This consequence could affect the decision to continue or not using the test, but it will not invalidate its interpretation and use.

2.3.6 Validation as Integration of Different Types of Evidence

> A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses (AERA et al., 2014, p. 21).

In order to determine what types of evidence should be collected to validate an interpretation of scores for an intended use of a test, it is useful to consider the propositions that support the intended interpretations and, subsequently, to look for relevant evidence.

Because multiple propositions can support a given use of a test, it is necessary to seek evidence for each one of them and integrate all the evidence to validate the interpretation. For example, we might want to judge the validity of the use of a science test for determining whether students are prepared to take a more advanced course in that subject.

A first statement that would support this use of the test is that the test scores effectively predict student success in the advanced science course. This statement would be examined with studies that provide evidence of relationships with other variables, such as the final grade in the advanced course.

But other arguments must also be considered, such as that the assessment procedures are equally suitable for all students taking the test. We should also make sure that there are no secondary variables that may have an impact on the test result. For instance, the use of higher level language may negatively impact on the results of examinees whose second language is English, which would not reflect a lack of scientific knowledge, but rather a lack of language proficiency (i.e., *a construct-irrelevant component*).

Another proposition that should be examined in this case is that all the knowledge areas to be measured in science, previously described in the specification of the content domain, are actually reflected in the test items, questions, and tasks (i.e., *adequate construct representation*).

Furthermore, we should look for evidence that the examinees' response process indeed reflects their knowledge in science and not, for example, their language and writing skills in open-ended questions and that the correction of the test is based on clear criteria not influenced by other irrelevant factors such as the examinees' handwriting (i.e., *construct-irrelevant components*).

Another proposition to evaluate would be that the different dimensions of the domain knowledge-in-science should be reflected in the internal structure of the test. Therefore, if the test evaluates the areas of biology, chemistry, and physics, then we would expect the relationships between the items to account for these three dimensions (i.e., *evidence on internal structure*).

Finally, another argument that should be examined is that the science test score is strongly and positively associated with scores obtained on other already accepted tests that measure knowledge in science (i.e., *evidence on relations to other variables*).

By gathering sufficient evidence of validity to support the interpretation of scores for an intended use of the test, a provisional validity judgment can be generated. This judgment may support the current interpretation and use of the test, or suggest redefinitions of the construct to be measured, changes in the test and its application, or may lead to future studies to investigate other areas of the test. In our example, it is possible that the first proposition is supported by the evidence, i.e., studies show that the science test score effectively predicts the final grade in the advanced course. However, in another study, we may observe that those examinees whose native language is not English, systematically obtain lower scores in the science test than the rest of the examinees, which could be associated to an irrelevant component of the construct to be measured. To examine this, we would require a measurement bias study (e.g., of differential prediction of the test between language groups) in order to establish whether this is a difference due to bias or due to effective differences between language groups. Then, we would have to look for explanations for this result and weigh the evidence, complement it with other studies, and come to a preliminary conclusion. It is important to note that integration of the different types of validity evidence is the most complex step of the entire validation process. This is the step that requires the strongest expertise in measurement and assessment. Although it is based on empirical evidence that is systematically reviewed, it also involves expert judgments that interact with the specific context and that should be made transparent by those responsible for validation.

New validity studies will be required each time the construct, method of assessment, application, or test population is modified. In these situations, further evidence of validity should be collected to support the interpretation of the scores in this new context. In summary, the validation process is a continuous process, which must be supported by the integration of all available evidence and should have an impact on the test itself and its use.

2.4 The Political Dimension of Validation

Undoubtedly, the political nature of educational measurement and assessment makes validation a political process. In what follows, we present numerous problems that stress the complexity of validation in a political context. We do this in order to generate awareness about this issue, since these political reasons are the ones that often create important obstacles to validation as a fundamental process to ensure the quality of educational assessment.

One key question that needs to be answered in every measurement and assessment system is: who is responsible for validation and who provides the necessary financial resources? Test developers are often contracted entities that depend on the perceived quality of their work. Therefore, they are less likely to critically evaluate the validity of their products themselves, unless they are explicitly asked to do so and to report on it, or they are highly professional and abide by the Standards as a basis for their professional self-definition.

Moreover, those who finance test development often do not have the technical capacity to understand how necessary it is for validation to take place; they follow a political logic and are more concerned with meeting deadlines, saving resources, and ensuring acceptance of the tests, while being less concerned with the assessment's quality. Finally, those affected by assessments often have even less technical expertise and sometimes less voice to effectively demand validation to support intended uses and avoid unintended uses of assessments.

Because of this, it is common that there is no clarity as to who should be responsible for validating a test. This entails that there are often no resources considered for test validation. However, the Standards are clear on this issue: test developers and those who make decisions based on test results are the ones responsible for conducting validation studies.

Now then, when should the validation take place? There are multiple conflicting interests regarding this question. Political agendas frequently demand that assessment programs are implemented quickly. Political commitment regarding assessment is often fragile, and thus the speed of assessment implementation is often paramount. Furthermore, validation is a complex process that must begin before the assessment system is put into action, but must also consider real-life implementation conditions in order to reach justifiable conclusions (particularly in contexts of high-stakes consequences). Finally, validation must be continuous, as evaluation contexts are subject to political pressures, and therefore change over time. In summary, validation should start as soon as possible after the decision to develop an assessment system has been taken, also considering real implementation conditions, and should then be a recurrent process that includes the changes occurring in the assessment context.

Taking into account political and time/resource pressures, is validation research independent enough to reach conclusions that may suggest changes to the assessment system in question? Reporting negative validation results is a sensitive issue, as it could be used as ammunition to end an assessment program, even though it would be unrealistic to expect to find only positive evidence regarding the validity of an assessment system. There is no easy or general solution to this problem, considering the specific context and interest groups, who generates the research agenda on validity, who decides how the validation results will be communicated, and who decides what modifications are implemented based on these results.

In relation to the implementation of changes based on validation, it would be important that the legal bases and regulations for assessment programs incorporate review cycles that allow these modifications to be made, especially in the case of high-stakes assessments. These legal foundations and regulations can provide continuity and useful guidelines for test developers and for those who make decisions based on test results. However, these guidelines must strike a good balance between being sufficiently explicit but not too detailed. The risk of rigid regulations is that validation research may not inform adaptations or modifications to the assessment method and implementation in a timely and feasible manner.

Furthermore, an essential condition for generating a research agenda on validity is that there are well-prepared and trained professionals. However, the capacity for validation research in Latin America and the Caribbean is still insufficient, and likewise, professionals in measurement are scarce. Hence, it is important to generate capacities in educational measurement and assessment in the region. Nonetheless, even outside of Latin America and the Caribbean, validation does not play the role that it should according to professional standards of good practice.

In sum, if the technical aspects of validation are complex, then the political dimension makes it even more intricate. This book is an important contribution to strengthening the awareness that in educational assessment it is indispensable to have standards of quality, particularly of validity, to ensure that decisions are supported by evidence. For evidence-based policy to be useful, educational assessments are required which results are used only for intended and validated interpretations and uses, and which unintended interpretations and uses are explicitly mentioned to promote their control.

"Validity theory is rich, but the practice of validation is often impoverished" (Brennan, 2006, p. 8). This conclusion is frequently expressed in educational assessment circles and, to our knowledge, remains applicable more than a decade later. However, there is less written about what can be done to overcome this division. Above all else, policy makers who implement educational assessment systems should be aware of the importance and complexity of arriving at valid inferences, and at the same time be mindful that a test or assessment system will always be subject to a certain degree of imprecision, uncertainty, and will be more useful to some users than to others. Also, those responsible for validating the proposed interpretations, uses, and consequences should be specified by regulation or law, while validation documentation should be made publicly available in a timely manner, before major decisions are made based on the data. Without the necessary formal or legal obligation, there may never be enough time and resources for validation to be carried out in a sufficiently rigorous manner. The evidence of validity can hardly be complete and definitive, but a serious effort should be visible.

In practice and in political circles, capacity in educational assessment is still lacking. Moreover, measurement experts often lack applied experience at the crossroads of policy and practice to be sufficiently sensitive to the political dimension of validation described previously. This contributes to the lack of mutual understanding and makes progress more difficult on the steep path to validation. We hope that this book will provide a basis for knowledge sharing and capacity building, and thus facilitate conversations about validity and validation among different stakeholders.

Returning to the title question, validation should not be a luxury, but a fundamental responsibility for educational assessment systems in Latin America and the world.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

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

American Psychological Association. (1952). Committee on test standards. Technical recommendations for psychological test and diagnostic techniques: A preliminary proposal. *American Psychologist, 7*(8), 461–475.

American Psychological Association, American Educational Research Association & National Council on Measurement in Education. (1954). Technical recommendations for psychological test and diagnostic techniques. *Psychological Bulletin, 51*(2), 201–238.

American Psychological Association, American Educational Research Association & National Council on Measurement in Education. (1966). *Standards for educational and psychological test and manuals*. American Psychological Association.

American Psychological Association, American Educational Research Association & National Council on Measurement in Education. (1974). *Standards for educational and psychological test*. American Psychological Association.

Borsboom, D., Mellenbergh, G., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 3–16). American Council on Education/Praeger.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods, 17*(1), 31–43. https://doi.org/10.1037/a0026975

Cronbach, L. J. (1949). *Essentials of psychological testing*. Harper & Brothers.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147–171). University of Illinois Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological test. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64). American Council on Education/Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kane, M. T. (2015). Explaining validity. *Assessment in Education: Principles, Policy & Practice, 23*(2), 198–211. https://doi.org/10.1080/0969594X.2015.1060192

Kane, M. T. (2016). Validity as the evaluation of the claims based on test scores. *Assessment in Education: Principles, Policy & Practice, 23*(2), 309–311. https://doi.org/10.1080/0969594X.2016.1156645

Kulik, J. A., Kulik, C. C., & Bangert-Drowns, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal, 20*(2), 435–447. https://doi.org/10.3102/00028312021002435

Lane, S., Parke, C. S., & Stone, C. A. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement: Issues and Practice, 17*(2), 24–28. https://doi.org/10.1111/j.1745-3992.1998.tb00830.x

Lane, S., & Stone, C. A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational Measurement: Issues and Practice, 21*(1), 23–30. https://doi.org/10.1111/j.1745-3992.2002.tb00082.x

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice, 16*(2), 14–16. https://doi.org/10.1111/j.1745-3992.1997.tb00587.x

Maguire, T., Hattie, J., & Haig, B. (1994). Construct validity and achievement assessment. *Alberta Journal of Educational Research, 40*(2), 109–126.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–100). American Council on Education.

Messick, S. (1994). Foundations of validity: Meaning and consequences in psychological assessment. *European Journal of Psychological Assessment, 10*(1), 1–9. https://doi.org/10.1002/j.2333-8504.1993.tb01562.x

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment *Educational Measurement: Issues and Practice, 14*(4), 5–8. https://doi.org/10.1111/j.1745-3992.1995.tb00881.x

Popham, W. J. (1997). Consequential validity: Right concern-wrong concept *Educational Measurement: Issues and Practice, 16*(2), 9–13. https://doi.org/10.1111/j.1745-3992.1997.tb00586.x

Ruch, G. M. (1924). *The improvement of the written examination.* Scott, Foresman and Company.

Santelices, V. (2021). Validity of assessment systems for admissions and certification. In J. Manzi, M. R. García, & S. Taut (Eds.), *Validity of educational assessment in Chile and Latin America* (pp. XX–XX). Springer.

Shepard, L. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5–8. https://doi.org/10.1111/j.1745-3992.1997.tb00585.x

Shepard, L. (2002). The hazards of high-stakes testing. *Issues in Science and Technology, 19*(2), 53–58.

Wiley, D. E. (1991). Test validity and invalidity reconsidered. In R. E. Snow & D. E. Wiley (Eds.), *Improving inquiry in the social sciences: A volume in honor of Lee J. Cronbach* (pp. 75–107). Erlbaum.

Zumbo, B. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R. Lissitz (Ed.), *The concept of validity* (pp. 65–82). Information Age Publishing.

Sandy Taut Deputy Head of Quality Agency, Bavarian State Office for Schools, Germany. Ph.D. in Education from the University of California, Los Angeles (UCLA), USA, and psychologist from University of Cologne, Germany. She has worked and researched issues related to educational assessment, teacher, instructional and school quality, and validation of measurement and assessment systems. Contact: staut@uc.cl

Siugmin Lay Psychologist from Pontificia Universidad Católica de Chile and Ph.D. in Psychology from Royal Holloway University of London, United Kingdom. She is currently an adjunct researcher at MIDE UC Measurement Center, Pontificia Universidad Católica de Chile. Her areas of interest are intergroup relations and attitudes studied from a social psychology perspective. Contact: splay@uc.cl