Giraldo, Frank
Validity and Classroom Language Testing: A Practical Approach 1
Colombian Applied Linguistics Journal, vol. 22, núm. 2, 2020, Julio-, pp. 194-206
Universidad Distrital Francisco José de Caldas
Bogotá, Colombia

Disponible en: http://www.redalyc.org/articulo.oa?id=305766351020

Validity and Classroom Language Testing: A Practical Approach[1]

La validez y la evaluación de lenguas en el aula de idiomas: un enfoque práctico

Frank Giraldo[2]

Abstract

Validity and validation are common in large-scale language testing. These topics are fundamental because they help stakeholders in testing systems make accurate interpretations of individuals' language ability and related ensuing decisions. However, there is limited information on validity and validation for classroom language testing, for which interpretations and decisions based on curriculum objectives are paramount, too. In this reflection article, I provide a critical account of these two issues as they are applied in large-scale testing. Next, I use this background to discuss and provide possible applications for classroom language education through a proposed approach for validating classroom language tests. The approach comprises the analyses of curriculum objectives, design of test specifications, analysis of test items, professional design of instruments, statistical calculations, cognitive validation and consequential analyses. I close the article with implications and recommendations for such endeavours and highlight why they are fundamental for high-quality language testing systems in classroom contexts.

Keywords: classroom language testing, language testing, validation, validity

Resumen

La validez y la validación son temas de discusión comunes en la evaluación de lenguas a gran escala. Estos temas son fundamentales porque permiten que aquellos involucrados en estos sistemas de evaluación puedan hacer interpretaciones claras, junto con las decisiones que de ellas se desprendan. No obstante, hay poca información en la literatura relacionada con la validez y la validación en contextos de aprendizaje de lenguas, donde las interpretaciones y decisiones basadas en objetivos curriculares también son fundamentales. En este artículo de reflexión, hago una revisión crítica de cómo estos dos temas son utilizados en evaluación a gran escala. Luego uso este contexto para discutir y presentar posibles aplicaciones para el aula de idiomas a través de una propuesta de enfoque para la validación de instrumentos de evaluación en este contexto. El enfoque incluye un análisis de objetivos curriculares, el diseño de especificaciones, el análisis de ítems en instrumentos de evaluación, el diseño profesional de evaluaciones, cálculos estadísticos, la validación cognitiva y, por último, análisis de consecuencias. El artículo lo concluyo con implicaciones y recomendaciones

---

1   This reflection article is on the validity of classroom language testing and connects theory and practice in validation.

2   Universidad de Caldas, Colombia. ORCID: https://orcid.org/0000-0001-5221-8245. frank.giraldo@ucaldas.edu.co

pertinentes para este proceso, además de enfatizar las razones por las cuales es vital para tener sistemas de evaluación de alta calidad.

Palabras clave: evaluación en el aula de clases, evaluación de lenguas extranjeras, validación, validez

Introduction

Validity is the most fundamental quality of testing systems, across social, professional and educational contexts. This assertion holds true whether tests are used in large-scale or classroom settings. Among assessment discussions, there is a consensus that tests are not valid: Validity is not the quality of an assessment instrument (e.g. a test) but relates to how appropriate interpretations are based on assessment data, for making particular decisions (Chapelle, 1999; Fulcher, 2010; Green, 2004; Messick, 1989; Popham, 2017). Thus, validity may be conceived as an abstract notion and an ideal. Because of the abstract nature of validity, validation has emerged as the data-gathering process to argue for the validity of interpretations and decisions made from tests. The quality and the process are crucial in large-scale and classroom language testing (Chapelle & Voss, 2013; Kane & Wools, 2019). Particularly, validation supports the development and monitoring of high-quality testing systems.

Validation research in language assessment abounds, specifically for large-scale testing—tests that affect many individuals (Bachman, 2004); such research is expected because of the consequences of using these instruments. Chapelle, Enright and Jamieson (2008) argue in favour of the validity of using the Test of English as a Foreign Language (TOEFL); the researchers claim that the TOEFL helps users make admission decisions for English-speaking universities that use academic English. Other examples of validation projects are assessments of the validity of using a placement test for international teaching assistants (Farnsworth, 2013), a web-based Spanish listening test to make placement decisions (Pardo-Ballester, 2010) and Llosa's (2007) comparison of a classroom test and a standardised test of English proficiency. These studies have collected data to claim the validity

of using these tests, used complex statistical calculations and compared these tests with other well-known instruments. Thus, validation research and discussions are predominant in assessing the validity of large-scale testing (Chapelle & Voss, 2013, Xi & Sawaki, 2017). However, the discussion on the validity and validation of classroom *language* testing has been limited, with researchers providing mostly a conceptual approach (see Bachman & Damböck, 2018; Chapelle & Voss, 2013; Kane, 2012).

Against this backdrop, the purpose of this reflection paper is twofold: to discuss validity as it relates to classroom *language* testing and *language* teachers and provide and reflect on strategies to validate classroom language tests such that they are manageable for teachers. I provide practical examples to demonstrate this process. I start the paper with an overview of definitions for validity and validation as central constructs and then discuss a practical approach for them in classroom language testing. I end the paper with implications of validating language tests, recommendations for validation and relevant limitations and conclusions.

Validity in Language Testing

Validity in language testing is about how logical and true interpretations and decisions are made based on scores (or in general data) from assessments. Validity has been considered a trait of tests: A test is valid if it measures what it has to measure and nothing more (Brown & Abeywickrama, 2010; Lado, 1961). However, this view is no longer used in educational measurement in general and in language testing specifically.

The following definition of validity in assessment is from the American Educational Research Association (AERA), American Psychological Association and National Council on Measurement in Education (NCME; 2014, p. 11): 'The degree to which evidence and theory support the interpretations of test scores for proposed uses of tests'. Earlier, Messick (1989, p. 13) provides a similar definition that since its inception was welcomed in language testing. To him, validity is 'an overall evaluative judgement of the degree to which evidence and theoretical rationales support

the adequacy and appropriateness of interpretations and actions based on test scores'.

Thus, in language testing, a score represents individuals' language ability and is used for making *decisions*, for example, to allow conditional admission to an English-speaking university (e.g. the aforementioned TOEFL case), or in a classroom, to move on to another unit in a course. This decision-making process is what Messick calls interpretations and actions, or uses of tests in AERA et al. (2014). The interpretations and actions should be appropriate because they are based on clearly defined constructs (i.e. language ability as a theoretical rationale) and on student performance on a test—what Messick and AERA et al. call evidence.

A couple of teachers using a placement test of reading comprehension with a group of new students at a language institute is an example of *evidence* and *theoretical rationale*. On the basis of the score from this instrument, a student is placed in Level II (decision or use). In this case, validity depends on demonstrating that 1) the student displayed a performance in reading that merited being in Level II (evidence) and 2) that the test was based on a clear definition of language ability for reading at Level II (theoretical rationale). If students start Level II and perceive that their skills are beyond those of their classmates, the interpretation (that the student had reading skills to be in Level II) and the decision (placing the student accordingly) are not valid. If the student is ready for Level II, there is validity in the interpretation and decision from this testing system.

To further explicate validity in language testing, the following hierarchy synthesises and simplifies this quality for the TOEFL (based on Chapelle et al., 2008). Tests serve purposes—they are not designed in a vacuum—and trigger the evidence (what test takers demonstrate) from which interpretations are derived. Subsequently, these interpretations are used to make claims and decisions about individuals.

Purpose: Measure a test taker's proficiency in academic English.
↓
Assessment of: Performance on the TOEFL (Evidence).

↓
Interpretations of: Test taker's state of academic English in listening, reading, speaking and writing (Theoretical Rationales).
↓
Claim: The student *does* or *does not* have sufficient academic English to study at university.
↓
Decision or use: Based on scores from the TOEFL, confer or deny conditional admission for university.

The aforementioned claim and decision must be validated; in other words, TOEFL developers must demonstrate through considerable amounts of research-based data that the claim and decision are valid, namely, logical and true. A similar approach can be used in classroom language assessment, in which the chain of logic as overviewed can be applied (see Bachman & Damböck, 2018; Chapelle & Voss, 2013; Kane, 2012). The following hierarchy is an example of a classroom language assessment for a listening quiz.

Purpose: Identify the students who are learning or having difficulty with listening skills A and B.
↓
Assessment of: Performance on a listening quiz with 20 multiple-choice questions; number of right and wrong answers (Evidence).
↓
Interpretations of: Students' level of listening comprehension as outlined in the course syllabus (Theoretical Rationale).
↓
Claim: The student who passes the quiz has the listening skills; the student who fails does not.
↓
Decision or use: If all students pass the quiz, they have developed the skills and are ready to develop new listening skills.

To argue for the validity of the aforementioned claim and decision, the teacher using this quiz must present evidence to demonstrate at least the following about the test:

• It is designed to activate skills A and B, and they are from the curriculum objectives.

- It was well designed to activate listening skills A and B.
- It *was not* designed to activate listening skills C and D.
- The students took the test without disruption; there were no problems with the administration.
- There were no instances of cheating.
- The teacher correctly checked the test and provided the relevant grades accurately: pass or fail.
- The answer key (the document that contains the correct answers) is accurate, namely, all the correct answers really are the correct answers.

To reiterate, validity is about how appropriate, logical and true interpretations and decisions are based on data from assessment instruments. If students cheated during this quiz, the score might be inflating their listening skills, the teacher is misinterpreting the data (correct answers) and those who passed may not really have the skills. Additionally, the decision to advance to other listening skills in the course is not valid. Notably, if the teacher mistakenly used a test for skills C and D, the interpretations and decisions are not valid, either. The test was not fit for purpose in this particular scenario.

Thus, validity for classroom testing can be likened to the definitions by AERA et al. (2014) and Messick (1989), with some modifications: Validity in classroom language testing depends on how appropriate interpretations and decisions are, based on the data from instruments used to activate the relevant language skills stated in a curriculum. As aforementioned, validity is an abstract concept. To make it practical, teachers can validate the tests they use for accurate interpretations and decisions, which I discuss next.

Validation in Language Testing

Validation is the process of evaluating the validity of a testing system. Validation entails the accumulation of empirical and theoretical evidence to demonstrate that a test has been used as expected and led to corresponding correct uses. Language testing professionals generally refer to validation as the process to estimate the validity of score-based interpretations, decisions and consequences (Bachman, 2005; Carr, 2011; Kane, 2006; Messick, 1994). Particularly, validation in large-scale testing requires the use of considerable amounts of quantitative and qualitative data (Xi & Sawaki, 2017), which in cases tend to be unnecessary for classroom testing (Brookhart, 2003; Popham, 2017). However, validation must also be acknowledged in classroom contexts because the validity of tests used in the classroom must be accounted for, too (Bonner, 2013; Brown & Hudson, 2002; Popham, 2017).

Specifically, I posit that validation in classroom language testing may help scrutinise the appropriateness of curriculum objectives, the overall quality of tests and the fairness with which students are treated in assessments. The validation schemes for classroom assessment reported in the literature (Bachman & Damböck, 2018; Bonner, 2013; Chapelle & Voss, 2013; Kane, 2012) have tended to be theoretical and offer general principles. However, according to my review of the literature, there are limited resources for language teachers to reflect and act upon the idea of validating the tests they use. Therefore, in the next section of this paper, I offer one possible praxis-based approach for examining the validity of interpretations and decisions as they emerge from using classroom language tests.

One Practical Approach for Validation in Classroom Language Testing

My proposed approach for validation in language classrooms comprises three major stages: The first stage relates to the congruence between curriculum objectives and the design of tests; the second stage is a close analysis of already-made instruments and the use of basic statistics; the last stage collects feedback to examine the consequences of using tests.

Curricular Focus

Scholars in educational measurement in general and those in language testing have argued that tests should reflect the skills, tasks, or content stipulated in a curriculum. This connection is

collectively called content validity (Bonner, 2013; Brown & Hudson, 2002; Douglas, 2010; Fulcher, 2010; Popham, 2017). If instruments collect evidence on students' stance against curriculum content, this evidence can be used to argue for the validity of an assessment.

Particularly, language teachers should ascertain whether the language skills from a syllabus are language related. For example, in Colombia, language learning is based on national standards stated in a document called *Guía 22* (Ministerio de Educación Nacional de Colombia, 2006, p. 22). Next, I present two examples that the document states as learning standards for *Reading* in English in sixth grade. I include a translation for each standard.

1) Puedo extraer información general y específica de un texto corto y escrito en un lenguaje sencillo.
I can extract general and specific information from a short text written in simple language.
2) Valoro la lectura como un hábito importante de enriquecimiento personal y académico.
I value reading as an important habit for personal and academic edification.

At face value, number 1) is a specific reading skill; however, number 2) is a skill that an individual can demonstrate regardless of language. Thus, 1) may be operationalised in a *language* test, namely, a teacher can create a reading quiz to assess the students' abilities in performing in this skill. Number 2) cannot be operationalised in a *language* test. Of course, the standards are meant to guide learning, teaching and assessment. The main point is that language teachers should observe how connected their language assessment instruments are to the skills of their language curriculum. Therefore, the main recommendation is for teachers to analyse whether the standards (or objectives) in their curriculum are language related, i.e. that they represent language ability. This notion is best encapsulated in this question: Can I design a test that provides me with information on my students' level/development of this learning standard (or competence) in the English language?

Test Specifications and Fit-to-Spec Analysis.

A practical approach for the curriculum level—and to have evidence for validity—relies on the design of test specifications, test specs for short (Davidson & Lynch, 2002; Fulcher, 2010). A document with specs describes how a test should be designed. Table 1 provides a simple example of a reading test.

Davidson and Lynch (2002) explain that teachers can conduct a fit-to-spec validity analysis. Once the 15 items for the test in Table 1 are designed, teachers can assess whether the items clearly align with the specs. To help teachers achieve this objective, I converted the descriptions in Table 1 into a checklist that teachers can use (Table 2).

Test specs and the results of fit-to-spec analysis are evidence for validation for three main reasons. First, the specs should naturally be based on the language skills stated in a syllabus, which can then provide evidence for the test's content validity. Second, the fit-to-spec analysis can unearth problematic items that are either assessing something not stated in the specs (and therefore not in the curriculum) or confusing the students. Finally, problematic items can be changed such that they better reflect the curriculum skills to be assessed. Appropriate specs and congruence between tests and curriculum objectives will most likely contribute to the validity of interpretations and, therefore, the purpose and decisions based on data.

Professional Test Design.

Another test development action in tandem with specs is the principled design of items and tasks. Language testing authors have provided guidelines for the professional design of tests (Alderson, Clapham, & Wall, 1995; Brown, 2011; Carr, 2011; Fulcher, 2010; Hughes, 2002). In particular, Giraldo (2019) synthesises ideas from these authors to provide checklists for the design of items and tasks. Table 3, which I adapted and modified from Giraldo (2019, p. 129-130), contains descriptors for a checklist that can be used to either design or evaluate a reading or listening test.

Table 1. Sample Test Specifications for a Reading Test

Purpose & Decision	The purpose of this test is to assess how students are developing the following reading skills. On the basis of the results from this test, the teacher and students can identify what they do well and what they must improve or reinforce before advancing to other reading skills.
Skills to be assessed	Identify the general message (moral) of tales. Identify specific details from narrative texts: characters' personalities, dates and places of events; and sequence of events (e.g. what occurs first and second).
Types & length of texts	1 fable. 1 classical tale (excerpt) 1 person's narrative account All texts are between 100 and 150 words.

Method & Instructions for design	Multiple-choice test with 15 questions
	3 options (A, B and C) for each question
	5 questions for the fable
	
	    Question 1 on the moral of the fable
	    Question 2 on an animal's character
	    Question 3 on a date
	    Question 4 on a place
	    Question 5 on a sequence of events
	
	5 questions for the classical tale
	    Question 6 on the main message of the tale
	    Question 7 on a character's personality
	    Question 8 on a date
	    Question 9 on a place
	    Question 10 on a sequence of events
	
	5 questions for the personal account
	    Question 11 on the main message of the account
	    Question 12 on the main character's personality
	    Question 13 on a date
	    Question 14 on a place
	    Question 15 on a sequence of events

Table 2. A Fit-to-Spec Analysis Table

Questions	Yes	No
Is Question 6 on the main message of the tale?		
Is Question 6 written for the students in the course?		
Is Question 10 on a sequence of events?		
Is Question 10 written for the students in the course?		
Please provide feedback on items 1 to 15:		

Table 3. Checklist of Guidelines for a Multiple-Choice Reading or Listening Test

Guidelines	Yes	No
The stem in Item # ___ is written clearly. *(If the stem is not clear for a fellow teacher or a student, it is probably not clear for the students with whom it will be used.)*		
The question in Item # __ **does not** have unknown vocabulary for students.		
All options in Item # __ are plausible, namely, they can be answered only by listening to/reading the text. *(If a student can guess the answer without listening or reading, the item is not assessing this construct.)*		
Item # __ **does not** provide the answer to another item. *(Item 4 may have information to answer Item 3. Check that this is not the case.)*		
Item # __ is independent of the other items. *(Each item in this test should be assessing one bit of the construct(s); thus, if items overlap, discard one of them.)*		
The correct answer (the key) for Item # __ **really** is the correct answer.		
Item # __ **only has one** correct answer. *(If the item has more than one answer, the options must be revised.)*		
Item #__is assessing one of the skills described in the test specs.		

A well-designed instrument should be a fundamental piece of evidence to argue for the validity of classroom language testing. A professional design helps strengthen the quality of assessments because they are constructed primarily to collect clear evidence on the constructs (i.e. skills) of interest, leading to accurate interpretations of and decisions on students' language ability. A poorly designed test might yield unclear information, undermining the overall validity of the assessment (Fulcher, 2010; Popham, 2003).

Statistical Calculations and Analyses.

Once the design of a test is complete and the instrument implemented, teachers may wish to conduct basic statistical analyses to evaluate their instruments, along with corresponding interpretations and decisions. Authors such as Bachman (2004), Brown (2011) and Carr (2008; 2011) have offered foundational explanations to calculate statistics for language testing. Excel, in Microsoft Office, can be used to perform calculations; the most important aspect is interpretations of the statistical data. Next, I propose simple calculations that can yield evidence for validation. I suggest that teachers use the calculations with which they feel most comfortable.

The context for the proposed statistics is a fictional diagnostic test of speaking. The example is that two teachers teaching the *Level III Speaking Skills Course* want to determine the speaking level of their 30 new students. To conduct this assessment, they use an interview format with a rubric that comprises these criteria: fluency, pronunciation, discourse management, grammar accuracy and vocabulary control.

The interview is based on the specific speaking skills for Level III; thus, the assumption is that the 30 students should *not* 'pass' the test (they are about to start Level III); in other words, the 30 students should not have the skills described in the rubric for this test. The passing score in this situation is 3.5.

- Calculate frequencies and percentages: The two teachers can observe what percentage of students were between these score ranges: 1.0 and 2.5, 2.6 and 3.4, and 3.5 and 5.0. Next, the teachers can interpret the percentages. For instance, if the score of 70% of students was between 3.5 and 5.0, they have the skills for Level III speaking and should be in another course. If the score of 70% of students is below this same score range (3.4 or lower), they are ready for the course. In both cases, an argument could

be that the diagnostic instrument yielded useful data to examine the validity of interpretations and decisions.

- Calculate mode, median and mean. The two teachers can observe the mode score, the median score and the mean score for all students. If the mode were 2.0, then the students with 2.0 are ready for Level III; if the median is 3.5, then 50% of students are ready for Level III and 50% seem to have the skills stated in the learning objectives for the course. Finally, if the mean (the average of all the 30 scores) is 4.0, the group has the speaking skills for Level III. Notably, high scores (5.0) may inflate the mean; thus, analysis of specific cases (e.g. low, failing scores) is warranted.

- Calculate mean and standard deviation. These two statistics are useful when analysed together. If the mean for the group of 30 students is 2.5 and the standard deviation (average distance of every score from the mean) is 0.2, then some students' score was 2.7 and others was 2.3. On the basis of this standard deviation, students are observed to have a similarly low level of speaking, interpreted as the group being ready for Level III. If the mean and standard deviation are 4.4 and 0.2, respectively, the group has the speaking skills for Level III. If the mean were 3.5 and the standard deviation for this particular test were 1.0, two phenomena are possible: The students have widely different levels of speaking, or there was little consistency in the assessment, as I explain next.

- Calculate the agreement coefficient and kappa for consistency. These two statistics help present the extent of the agreement between two test administrations, two raters, or two score-based decisions such as *pass* and *fail*. In the aforementioned diagnostic test example, suppose the two teachers assessed each student at the same time, so each student received two scores. If the agreement coefficient is 70%, the two teachers made the same decisions (pass or fail) in 70% of the cases (21 students). The performance of the other 30% (9 students) needs to be revised. If kappa, a detailed calculation

for consistency, is 85%, the agreement level between the two teachers is very high (Fulcher, 2010). Consistency in this scenario can be interpreted as the two teachers using the rubric accurately: They understood the constructs (e.g. grammar accuracy, fluency) and assessed them fairly while they heard students speaking during the interview.

- Calculate means and standard deviations in a differential groups study (Brown & Hudson, 2002). This type of study requires a somewhat higher level of sophistication than the previous calculations. The two teachers can use the same interview and corresponding rubric with students who are in the *Level IV Speaking Course* and compare their performance with the means and standard deviations of the students about to start Level III. The assumption in this case is that students in Level IV should pass the interview because they have the skills presented in Level III: The mean should be high and the standard deviation low. Both the mean and standard deviation for the students about to start Level III should be low. If a high percentage of students in Level IV fail the diagnostic interview for Level III, the instrument must be investigated, and the validity of inferences and decisions from it must be questioned. Perhaps determining what occurred during the Level III course is necessary.

The statistical calculations in the aforementioned speaking scenario provide information on students' speaking skills vis-à-vis the Level III course. For validation purposes in general, statistics can be used to argue for the validity (or lack thereof) of language tests. For example, if in the aforementioned testing scenario kappa is low (20% or less), the two teachers disagreed widely and, therefore, interpretations and decisions cannot be trusted –they are not valid. The central point is that for statistics to help with validation, they must be interpreted against the constructs and the purposes for which a test is used.

Cognitive Validation

Authors such as Bonner (2013) and Green (2014) have suggested that teachers ask students for

insights into assessment processes and instruments or observe students as they take tests. The idea of cognitive validation is to stimulate students' thinking and reflection regarding language assessment. Bonner, for example, recommends the use of think-alouds, observations and interview protocols to tap into students' cognition. For example, teachers can ask students the following questions (in an oral interview or written open survey) to collect evidence for the validity of interpretations and decisions:

1. How did you feel while [writing your narrative text]?
2. What skills do you feel the [narrative task] was assessing? Do you feel you had the opportunity to demonstrate these skills on this test?
3. If anything, what was difficult for you in this [narrative task]?

For ease of use, the three questions can be asked in the language with which students are most comfortable. The answers can then be used to investigate the validity of a given instrument. For instance, if a student feels the instructions for a task were difficult to understand, and the teacher notices that his/her performance was poor, maybe the instructions caused the poor performance. In this case, interpretations and decisions must be challenged and studied carefully. If students report that the instructions were clear and they performed well, this piece of evidence supports the validity of interpretations and decisions. Similarly, if students' answers to question 1 reflect what test specs stipulate, this observation can also be used as evidence.

Analysis of Consequences.

Generally, assessments should lead to beneficial consequences, especially when assessments are used for instructional purposes (Bachman & Damböck, 2018; Green, 2014; Kane & Wools, 2019). By and large, the consequence of classroom language testing should be improved language learning. Thus, a final proposed action for validating classroom language tests is to analyse their consequences. In Table 4 is a list of categories related to purposes for classroom language testing, with proposed courses of action.

As Kane and Wools (2019) reiterate, classroom assessments should be useful in attaining instructional purposes and their validity assessed on the extent to which these objectives are fulfilled. The proposed questions for a consequential analysis in Table 4 might help teachers evaluate the reach and usefulness of their tests.

The steps in the proposed practical approach for validating classroom language tests, summarised in

Table 4. Language Testing Purposes and Analysis of Consequences

Purposes	Consequential Analysis
Diagnostic	After providing feedback on the diagnostic, ask students and teachers in the corresponding courses how students are feeling/doing. For example: If the diagnosis stated that the student needed to be in the course, she/he should feel fine in it. Is she/he improving language?
Progress	If after a progress test, students require additional emphasis on a particular language skill, provide the necessary review/reinforcement tasks and ask students whether the tasks are helping them with the areas that need attention.
Achievement	For students who failed the test and had to repeat the course: To what extent are you now improving the language skills for this course? For students who passed the test and are now in a new course: To what extent do you feel prepared for this course? Are you doing well? Do you feel you learned the skills/contents from the last course? To the teacher: To what extent do you feel these students are prepared for this course? Are they doing well? Do you feel students achieved the learning objectives from the last course?

Figure 1, have language constructs as a common factor. Whether language teachers assess language ability, reading, speaking, or any other language skill, validity and validation in the language classroom use constructs as a central notion. First, language tests collect evidence on language curriculum objectives operationalised through test specs; second, a fit-to-spec analysis is concerned with the quality of items and tasks in relation to the specs; third, a professional design ensures that the correct constructs can be triggered through the correct means (i.e. instruments); fourth, statistics can be useful especially when interpretations help teachers analyse constructs; finally, cognitive validation and consequential analysis engage stakeholders in discussing, from a qualitative perspective, the constructs and appropriateness of instruments. Together, the evidence from these steps can be used to gauge the validity of classroom language tests: The relative accuracy of interpretations and decisions from classroom test data.

Implications and Recommendations

Validating a classroom language test may imply the use of documentary evidence (e.g. from test specs) and empirical data (e.g. percentages from a statistical calculation) to support validity. Such endeavours may also entail a considerable amount of work that may be too much for language teachers to perform. In such a case, I suggest that validation be performed for tests that are formal (e.g. an achievement test), for which the stakes are high and consequences impactful for students. On an everyday basis, such as when teachers conduct an informal, alternative assessment, they may only be concerned with how assessment data is feeding back on teaching and learning and representing instructional goals (Kane & Wools, 2019).

Another implication of validation for classroom testing that might emerge from the steps in this paper is the need for language assessment
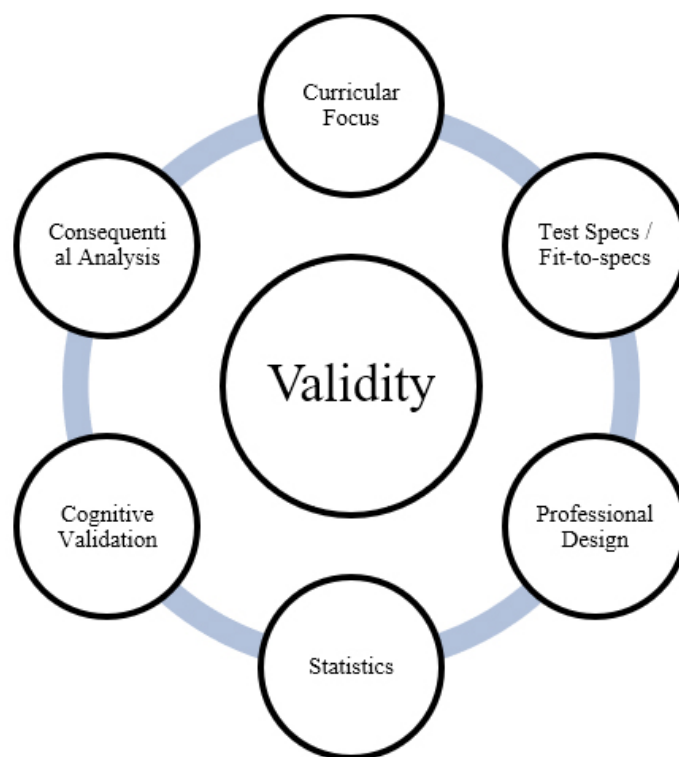

Figure 1. Sources of Evidence for Validity in Classroom Language Testing

literacy –LAL– (Fulcher, 2012; Inbar-Lourie, 2017). In other words, teachers may need satisfactory understanding of theoretical knowledge and skills for language testing, dimensions understudied in language education programmes (Giraldo, 2018; Herrera & Macías, 2015; López & Bernal, 2009; Vogt & Tsagari, 2014). For example, teachers must know how to calculate and, most importantly, interpret statistical information to evaluate validity in testing. As a recommendation for promoting LAL, teachers may use language testing textbooks or online resources; some of these are open source, for example, TALE Project (Tsagari et al, 2018), which includes a handbook to study language assessment issues.

Limitations

In this article, I propose one approach to validating language tests in the classroom. Thus, this is not an all-encompassing treaty; as authors in validation research have expressed, approaches to collecting evidence for validity considerations can vary widely (Chapelle, 1999; 2012; Kane, 2012). A specific approach will most necessarily depend on the particular purposes, characteristics and needs of a given educational context. As aforementioned, LAL might be necessary for validation; thus, the higher the LAL of stakeholders, the more robust a validation study can be.

Another limitation in this paper, primarily due to space constraints, is the validation of alternative assessment systems. My reflections and discussion in this paper leaned toward a summative view of testing because, as explained in the implications section, formal tests should be validated more systematically given the consequences they entail. Thus, validation for alternative schemes in assessment may warrant further study, which I predict will resort to qualitative research.

A related limitation refers to the use of task-based assessment in the classroom (Norris, 2016). The discussions in this paper covered general language courses in which language ability is the overarching construct. Conversely, in task-based scenarios, stakeholders may be more interested

in observing what real-life tasks individuals can perform using language (Long, 2015). Thus, in classrooms where task-based language assessment is the guiding methodology, other approaches to validation are warranted.

Finally, a limitation of the validation approach I discuss is that statistical analyses may not be a common topic for language teachers and may require further LAL, as aforementioned. As I state in this paper, validation is about collecting evidence from various sources, and statistics is one source. Language teachers attempting to validate classroom tests should, ultimately, analyse their expertise for their validation schemes for a given test and related purpose. The present proposal may be a guide for where to start their validity endeavour.

Conclusions

Validity and validation should be concerns in high-quality classroom language testing, and their relevance should not be limited to large-scale testing. Students, teachers and educational systems are the direct recipients of language tests. Thus, the purpose of this paper was to reflect on validity and validation as necessary discussions for language teachers, along with one possible practical approach to validation. Fundamentally, validity in classroom language testing reflects the relative appropriateness and accuracy of interpretations and decisions based on data from instruments which hopefully trigger instructional objectives for language learning. Validation in this scenario involves collecting evidence from various sources to evaluate the validity of interpretations and decisions in classroom language testing.

The approach I propose in this paper includes three stages: formulating curriculum objectives and specifications; designing and analysing the test items and tasks and the data from tests; and using a qualitative, student-based methodology. As aforementioned, this is *one* proposed approach; thus, teachers may be interested in studying and experimenting with different forms in which validation can be conducted. I posit that case studies of teachers validating their classroom language tests

may advance the field of language testing. These reports can contribute to the width and breadth of validation in language education. The goal of such an enterprise must be consolidating assessment systems that are valid and useful for supporting language learning.

References

Alderson, C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

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

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge University Press.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2(1), 1-34. https://doi.org/10.1207/s15434311laq0201_1

Bachman, L. F., & Damböck, B. (2018). *Language assessment for classroom teachers*. Oxford University Press.

Bonner, S. M. (2013). Validity in classroom assessment: Purposes, properties, and principles. In J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 87–106). Sage.

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice, 22*(4), 5–12. https://doi.org/10.1111/j.1745-3992.2003.tb00139.x

Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practice*. Pearson Longman.

Brown, J. D. (2011). *Testing in language programs: A comprehensive guide to English language assessment*. McGraw Hill.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge University Press.

Carr, N. T. (2008). Using Microsoft Excel® to calculate descriptive statistics and create graphs.

Language Assessment Quarterly, 5(1), 43-62. https://doi.org/10.1080/15434300701776336

Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford University Press.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, 19, 254-272. https://doi.org/10.1017/S0267190599190135

Chapelle, C. A., Enright, M. K., Jamieson, J. (Eds.) (2008). *Building a validity argument for the test of English as a foreign language*. Routledge.

Chapelle, C. A. (2012). Conceptions of validity. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 21-33). Routledge.

Chapelle, C. A., & Voss, E. (2013). Evaluation of language tests through validation research. In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (pp. 1079–1097). John Wiley and Sons, Inc.

Davidson, F., & Lynch, B. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. Yale University Press.

Douglas, D. (2010). *Understanding language testing*. Routledge.

Farnsworth, T. (2013). An investigation into the validity of the TOEFL iBT speaking test for international teaching assistant certification. *Language Assessment Quarterly, 10*(3), 274-291. https://doi.org/10.1080/15434303.2013.769548

Fulcher, G. (2010). *Practical language testing*. Hodder Education.

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2), 113-132. https://doi.org/10.1080/15434303.2011.642041

Giraldo, F. (2018). Language assessment literacy: Implications for language teachers. *Profile: Issues in Teachers' Professional Development, 20*(1), 179-195. https://doi.org/10.15446/profile.v20n1.62089

Giraldo, F. (2019). Designing language assessments in context: Theoretical, technical, and institutional considerations. *HOW Journal, 26*(2), 123-143. https://doi.org/10.19183/how.26.2.512

Green, A. (2014). *Exploring language assessment and testing*. New York, USA: Routledge.

Herrera, L., & Macías, D. (2015). A call for language assessment literacy in the education and development of Teachers of English as a foreign language. *Colombian Applied Linguistics Journal, 17*(2), 302-312. https://doi.org/10.14483/udistrital.jour.calj.2015.2.a09

Hughes, A. (2002). *Testing for language teachers: Second edition*. Cambridge University Press.

Inbar-Lourie, O. (2017). Language assessment literacy. In E. Shohamy, S. May, & I. Or (Eds.), *Language Testing and Assessment* (3rd ed., pp. 257-268). Springer.

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

Kane, M. (2012). Articulating a validity argument. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 34-47). Routledge.

Kane, M., & Wools, S. (2019). Perspectives on the validity of classroom assessments. In S. Brookhart & J. McMillan (Eds.), *Classroom assessment and educational measurement* (pp. 11-26). Routledge.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. McGraw Hill.

Llosa, L. (2007). Validating a standards-based classroom assessment of English proficiency: A multitrait-multimethod approach. *Language Testing, 24*(4), 489-515. https://doi.org/10.1177/0265532207080770

Long, M. (2015). *Second language acquisition and task-based language teaching*. John Wiley and Sons, Inc.

López, A., & Bernal, R. (2009). Language testing in Colombia: A call for more teacher education and teacher training in language assessment. *Profile: Issues in Teachers' Professional Development, 11*(2), 55-70.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*. (3rd ed., pp. 13-103). Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*, 13–23. https://doi.org/10.3102/0013189X023002013

Ministerio de Educación Nacional de Colombia (2006). *Estándares básicos de competencias en lenguas extranjeras: Inglés. Formar en lenguas extranjeras: ¡el reto! Lo que necesitamos saber y saber hacer*. Imprenta Nacional.

Norris, J. (2016). Current uses for task-based language assessment. *Annual Review of Applied Linguistics, 36*, 230–244. https://doi.org/10.1017/S0267190516000027

Pardo-Ballester, C. (2010). The validity argument of a web-based Spanish listening exam: Test usefulness evaluation. *Language Assessment Quarterly, 7*(2), 137-159. https://doi.org/10.1080/15434301003664188

Popham, J. (2003). Test better, teach better. *The instructional role of assessment*. Association for Supervision and Curriculum Development.

Popham, J. (2017). *Classroom assessment: What teachers need to know*. Eighth edition. Pearson.

Tsagari, D., Vogt, K., Froelich, V., Csépes, I., Fekete, A., Green A., Hamp-Lyons, L., Sifakis, N. &, Kordia, S. (2018). Handbook of assessment for language teachers. Retrieved from: http://taleproject.eu/

Vogt, K., & Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly, 11*(4), 374-402. https://doi.org/10.1080/15434303.2014.960046

Xi, X., & Sawaki, Y. (2017). Methods of test validation. In E. Shohamy, I. G. Or, & S. May (Eds.), Language testing and assessment: Encyclopedia of language and education (3rd ed., pp. 193-210). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-02261-1_19

206

Giraldo, F. (2020) • Colomb. Appl. Linguist. J.
Printed ISSN 0123-4641 Online ISSN 2248-7085 • July - December 2020. Vol. 22 • Número 2 pp. 194-206.