FOLIOS

Folios
ISSN: 0123-4870
ISSN: 0120-2146
Universidad Pedagógica Nacional

Giraldo, Frank; Naranjo-Trujillo, Dira Estefania; Ariza-Villa, Jairo Alonso
From the Design of Assessments to Language Assessment Literacy
Folios, no. 58, 2023, July-December, pp. 126-139
Universidad Pedagógica Nacional

DOI: https://doi.org/10.17227/folios.58-16385

Available in: https://www.redalyc.org/articulo.oa?id=345977459009

FOLIOS

Segundo semestre de 2023 • pp. 126-139
Segunda época
N.º 58

From the Design of Assessments to Language Assessment Literacy

Do desenho de avaliações ao letramento na avaliação de línguas

Desde el diseño de evaluaciones y hacia la literacidad en evaluación de lenguas

Frank Giraldo*
Dira Estefania Naranjo-Trujillo**
Jairo Alonso Ariza-Villa***

Para citar este artículo

Giraldo, F., Naranjo-Trujillo, D. E. y Ariza-Villa, J. A. (2023). From the Design of Assessments to Language Assessment Literacy. *Folios,* (58), 126-139. https://doi.org/10.17227/folios.58-16385

* Professor and Researcher at the Foreign Languages Department of Universidad de Caldas in Manizales, Colombia
Correo electrónico: frank.giraldo@ucaldas.edu.co
** Student in the BA in Modern Languages of Universidad de Caldas, Manizales, Colombia.
Correo electrónico: dira.221710692@ucaldas.edu.co
*** Student in the BA in Modern Languages of Universidad de Caldas, Manizales, Colombia.
Correo electrónico: jairo.221713426@ucaldas.edu.co

Artículo recibido
26•03•2022

Artículo aprobado
13•02•2023

Segunda época
N.º 58
Segundo semestre de 2023 • pp. 126-139

Abstract

Language Assessment Literacy (LAL) is now a core topic of discussion in language testing. Research has indicated that language teachers are interested in improving this area in their professional development. This research paper reports on an evaluation of English teachers' assessment instruments through qualitative content analysis. We analyzed 60 assessment instruments, 30 for receptive skills (reading or listening) and 30 for productive skills (speaking or writing). In the corpus, we found items and tasks that had both technical and theoretical issues. Based on our findings, and to foster teachers' LAL, we conclude that a program for these stakeholders should focus on the careful design of assessment instruments and the study of qualities of language assessments in context.

Keywords
assessment design; assessment instruments; language assessment literacy; language assessment qualities; language skills

Resumo

O letramento na avaliação de língua estrangeira (LAL) é hoje um tema central de discussão na avaliação de idiomas. Várias investigações têm indicado que os professores de línguas estão interessados em aprimorar essa área em seu desenvolvimento profissional. Este artigo de pesquisa examina instrumentos de avaliação elaborados por professores de inglês, por meio da análise de conteúdo qualitativa. Foram analisados 60 instrumentos de avaliação, sendo 30 de habilidades receptivas (ler ou ouvir) e 30 de habilidades produtivas (falar ou escrever). No corpus, encontramos itens e tarefas que apresentavam problemas técnicos e teóricos. Com base em nossas descobertas, e para promover o LAL dos professores, concluímos que um programa para as partes interessadas deve se concentrar no desenho cuidadoso de instrumentos de avaliação e no estudo das qualidades de avaliação de língua estrangeira em contexto.

Palavras-chave
desenho de avaliação; instrumentos de avaliação; letramento na avaliação de línguas; avaliação das competências linguísticas; habilidades linguísticas

Resumen

La literacidad en la evaluación de lenguas extranjeras (LEL) es ahora un tema central de discusión en la evaluación de idiomas. Diversas investigaciones han indicado que los profesores de idiomas están interesados en mejorar esta área en su desarrollo profesional. Este artículo de investigación examina los instrumentos de evaluación diseñados por profesores de inglés, a través del análisis de contenido cualitativo. Se analizaron 60 instrumentos de evaluación, 30 de habilidades receptivas (lectura o escucha) y 30 de habilidades productivas (habla o escritura). En el corpus, encontramos ítems y tareas que tenían problemas tanto técnicos como teóricos. Con base en nuestros hallazgos, y para fomentar la LEL de los profesores, concluimos que un programa para las partes interesadas debe centrarse en el diseño cuidadoso de los instrumentos de evaluación y el estudio de las cualidades de la evaluación de lenguas extranjeras en contexto.

Palabras clave
diseño de evaluaciones; instrumentos de evaluación; literacidad en la evaluación de lenguas; evaluación de habilidades lingüísticas; habilidades lingüísticas

Introduction

The field of language testing and assessment has been increasingly concerned with the need to foster teachers' Language Assessment Literacy (LAL): the knowledge, skills, and principles necessary for sound assessment in context (Davies, 2008; Inbar-Lourie, 2008, 2012). Thus, LAL has become a core dimension of the language teaching profession. In the existing research, teachers have reported that they want to learn about assessment, especially from a practical perspective, which is the *skills* component of LAL (Brindley, 2001; Fulcher, 2012; Malone, 2017).

In LAL, *knowledge* refers to theories of language ability, language learning, acquisition, and assessment; as well as frameworks for assessment; particular assessment contexts, and others (Davies, 2008; Inbar-Lourie, 2013). *Principles* include ethical and fair uses and practices of assessment and critical approaches scrutinizing the impact of assessment (Davies, 2008; Giraldo, 2018a). *Skills*, which teachers seem most interested in, involve the development of assessment instruments for traditional and alternative uses; ability to connect assessment with teaching and learning; providing clear feedback on student performance; technologies for assessment; and statistics for score interpretation (Fulcher, 2012; Giraldo, 2018a; Malone, 2017; Taylor, 2013).

Teachers' need to develop skills in language assessment is sensible. They resort to assessment instruments to account for language learning, and poorly designed assessments may be detrimental, as they may not collect relevant information to account for such learning. In the existing literature, there is scarce information about analyzing language teachers' assessment instruments as a window into their LAL and as a basis for fostering this dimension of their teaching practice (as examples, see Frodden et al., 2004; Giraldo, 2018b; and Levi & Inbar-Lourie, 2019). Thus, in the present study, we report on the findings from a research study which analyzed the assessment instruments designed by a group of state high school English language teachers in Colombia.[1] The research questions that guided this study were as follows:

> What are the characteristics of a set of language assessments designed by a group of English language teachers?
>
> How can these instruments be used to plan an online assessment course for these stakeholders?

The findings allowed us to describe these teachers' skills and areas for improvement in test design and to derive implications to plan an online course in language assessment. We start this report with a theoretical central overview and technical considerations for designing useful language assessments; also, we explore existing research in the area, and then, we discuss why professional design of assessments is paramount. Later, we explain our research methodology and the findings, followed by a discussion of how they are useful for planning LAL courses. We close the paper with limitations and recommendations for LAL courses elsewhere.

Theoretical Framework

Characteristics of Useful Language Assessments

An assessment is an instrument in which students can show whether they have learned or progressed in their language skills. Examples of assessment instruments include traditional ones, such as tests with multiple-choice, true-false, and matching items; and rubrics for assessing speaking or writing performance. Alternative instruments include self-assessment or peer-assessment checklists and protocols for managing portfolios. Whatever the type, the educational purpose of language assessment is to collect clear information about either skills, contents, or learning objectives in a particular language curriculum (Brown, 2011;

1 The present report is part of a larger research study, which sought to describe and foster the LAL of thirty English language teachers. In this report, we only focus on the analysis of assessment instruments as a source to draw teachers' LAL, particularly design skills.

Green, 2014). Then, the information collected through these means is used to document or further foster language learning and teaching (Green, 2014; Bachman & Damböck, 2018) within a particular language education context.

To further contribute to an assessment's usefulness, teachers designing or using assessment instruments should consider their theoretical qualities. Table 1 is a synthesis of five central qualities of assessments (Bachman & Palmer, 2010; Giraldo, 2019; Green, 2014):

Table 1. *Qualities of Language Assessments*

Quality	Definition
Validity	The degree to which interpretations of performance data, collected through an assessment, are appropriate for intended purposes. For this, the assessment should tap into the particular language skills it is designed to activate.
Reliability	The degree to which there is consistency in scoring performance, e.g., in a speaking assessment based on a rubric used by two raters. The degree to which a closed-ended test (e.g., true-false) yields similar results if used twice with the same group of students.
Authenticity	The degree to which the assessment situation, and the language used in it, are similar to how people use language in real-life situations.
Practicality	The degree to which there are enough resources to make the development of an assessment viable.
Washback	The degree to which an assessment influences teaching and learning in the context where the assessment is used; this influence can be positive or negative.

Source: Author's own elaboration (2021).

All the qualities above are relative; hence, the use of the phrase *the degree to which*: an assessment is not totally authentic or inauthentic, but it largely depends on the purpose and context where it is used, as it is commented above. One final feature of a useful assessment is that it is professionally designed —following technical design guidelines — to achieve its purpose. Below we review core guidelines for creating language assessments.

Technical Guidelines for Designing Language Assessments

For quality control, assessment designers should plan their products by writing a document with specifications (Davidson & Lynch, 2001; Fulcher, 2010). This document includes the purpose, the specific skills, the assessment method, number of items or tasks and their nature, as well as any other information that can help to ensure the assessment is planned and designed to meet its purpose. Among these specifications, one needs to be underscored: the specific skills to be assessed need to be clearly delineated; this is commonly called construct definition (Carr, 2011; Fulcher, 2010). Lack of clarity in this regard may lead to an assessment that is not useful. In Table 2, we synthesize major specific guidelines for listening/reading and speaking/writing assessments.

Table 2. *General Guidelines for Designing Instruments to Assess Language Skills*

Listening and reading	Speaking and writing
- The items (multiple-choice, true-false, matching, open questions, and others) for the assessment should be based on the instructions laid out in the specifications document. - The items should be written for the particular proficiency level of a group of students. - All the items in the test should have the potential to activate and be aligned with the specific construct for the assessment. - The items in the assessment should be answered only by listening/reading and not through guessing, for example. - The instructions in the assessment should be clear and succinct so the focus is on reading/ listening comprehension of the source text. - The items should activate understanding of ideas in the context of the source text, rather than linguistic knowledge, e.g., a grammar rule.	- The assessment task (situation, rubric, interaction) should be based on the specifications stated for the assessment. - The assessment task involves performance, i.e., it activates how students use speaking or writing in communicative scenarios; in other words, the task does not assess explicit linguistic knowledge. - The rubric for assessing student performance is construct-relevant (i.e., it assesses speaking/writing skills). - Each sub-skill in the rubric (grammar, pronunciation, punctuation, etc.) needs to have clearly written descriptors that explain the given performance that is necessary for task completion. - Users of the rubric should be trained to use it and find it clear. - The assessment task involves realistic language-driven purposes, e.g., asking for a favor, inquiring for information; informing, etc.

Source: Author's own elaboration (2021).

As stated, the previous table synthesizes fundamental guidelines for test construction; nevertheless, for design specifics, other specialized works can be consulted: Buck (2001) for listening; Alderson (2010) for reading; Luoma (2004) for speaking; and Cushing (2010) for writing. For test construction in general, useful resources are Alderson et al. (1995) and Carr (2011).

Related Research

As we mention earlier, limited research has been done regarding analyses of language assessment instruments. The first three studies below aimed to describe language teachers' assessment instruments used in their assessment practices. The second set of studies described the assessment instruments teachers designed as they were engaged in language assessment courses.

Frodden et al. (2004) studied the assessment instruments used by foreign language teachers (English and French) working in two Colombian universities. In their findings, the authors report that teachers placed emphasis on assessing vocabulary and grammar, but not so much on authentic language use. Additionally, the instruments tended to have problems with construct validity, i.e., it was not clear what they really aimed to assess as, for example, no scoring procedures were stated. Finally, as the authors report, traditional instruments were used more so than were alternative assessments.

Giraldo (2018b) conducted a study delving into English teachers' beliefs and practices involving the design and use of final achievement tests. The instrument analyses indicated that teachers in this study assessed linguistic and pragmatic aspects of language, with minor attention to sociolinguistic issues. Additionally, while tests tended to assess language in context, there were problematic areas in rubric design that led to reliability issues, i.e., lack of clarity regarding how to score language skills.

In a recent study, Villa-Larenas and Brunfaut (2022) critically examined the LAL of twenty language teacher educators in Chile. As part of their analysis, the researchers studied a set of assessment instruments used by these stakeholders. The authors indicate that the teacher educators used a variety of assessment techniques, most notably "fill-in-the-gaps, […] constructed response, matching, sequencing, and sentence completion" (p. 11).

The existing literature indicates the positive impact of language assessment training on teachers' design of assessment instruments. Overall, authors report that teachers design authentic language use tasks connected to their classrooms (Koh et al., 2018; Montee et al., 2013); clearly operationalize the language skills to be assessed (Arias et al., 2012; Koh et al, 2018). Additionally, these authors report that the programs were based on contextual needs analysis of language teachers' LAL, which included the study of assessment instruments. Thus, the research has suggested that analyzing assessment instruments may be used as one source of feedback for planning and implementing successful LAL training for teachers.

Why does the Design of Assessments Matter?

From the language performance that students show in an assessment, teachers are supposed to determine whether students have developed language skills; also, the instrument is designed to meet a purpose. Thus, properly designed assessments aid in doing the aforementioned tasks; poorly designed assessments may give erroneous results about students' language skills and, therefore, not be fit for a given purpose.

Another reason why design in assessment is crucial is the LAL needs teachers have reported in this regard. Although teachers design or resort to already designed instruments, studies have shown that they want and expect training in how to professionally construct instruments (Fulcher, 2012; Hasselgreen et al., 2004; Vogt & Tsagari, 2014). In response to this need, LAL programs have indicated that a design-based course impacts teachers' LAL: primarily knowledge and skills, with secondary attention to principles (Giraldo, 2021).

Against this background, in this paper we report an analysis of language assessment instruments in order to elucidate and interpret the *skills* dimension (particularly instrument design) of a group of LAL teachers. This analysis helped us to generate implications and recommendations for planning the language assessment course for the teachers in our study. We provide details from our research in the methodology and findings sections below.

Methodology

Our study was grounded on a qualitative approach because we wanted to interpret educational phenomena, namely a particular component of the LAL of English language teachers through the analysis of documents, i.e., assessments they designed. For scrutiny, we used Schreier's (2012) qualitative content analysis approach. This methodology relies on a coding scheme with two perspectives: on the one hand, instruments were analyzed conceptually; this means we looked for trends in the instruments and related them to theoretical and technical concepts in language assessment, e.g., authenticity and guidelines for item design (see the two previous sections); on the other hand, we used a complementary approach that was guided by data, searching for design trends in the *corpus* of instruments to identify strengths and aspects to improve in design.

Participants

Thirty English language teachers consented to share two assessment instruments, which formed the *corpus* in our study. These teachers agreed to participate in an online language assessment course. To design the course, a major source to draw these participants' LAL were the instruments we studied. However, we also asked them about their LAL through a questionnaire and an individual interview.

In terms of the participants, fifteen teachers had an MA degree, fourteen a BA, and one a specialization, as their highest educational level. Their teaching experience is presented in Table 3, while some of their assessment practices can be seen in Table 4, both with rounded percentages.

Table 3. *Experience Teaching English*

Time	n	%
Less than 1 year		
1-5 years	7	24%
6-10 years	10	33%
11-15 years	5	17%
16-20 years	5	17%
21-25 years	1	3%
26-30 years	1	3%
More than 30 years	1	3%

Source: Author's own elaboration (2021).

Table 4. *Selected Teachers' Assessment Practices*

Practices	Yes	No
I design assessments with multiple choice questions.	30 (100%)	
I design assessments with True and False statements.	29 (97%)	1 (3%)
I design rubrics.	21 (70%)	9 (30%)
I explicitly align the assessments I use to the objectives of the courses I teach.	23 (78%)	7 (22%)
I evaluate the assessments instruments used in class (i.e., I check whether they have a good quality).	23 (78%)	7 (22%)
I assess other aspects besides English.		
Effort	30 (100%)	
Discipline	27 (90%)	3 (10%)
Punctuality	26 (87%)	4 (13%)
Attendance	22 (73%)	8 (27%)
Responsibility	30 (100%)	
Respect	28 (93%)	2 (7%)

Source: Author's own elaboration (2021).

The Corpus

The *corpus* consisted of 60 language assessment instruments. From this *corpus*, we analyzed 51 and excluded 9 because they were lesson plans and not assessments. The final 51 instruments were divided into the four major language skills: listening, reading, speaking, and writing. The choice of assessments for these skills is derived from the standards for language learning in Colombia, which state that these skills are part of the language curriculum for high schools. In the Colombian educational context, English language teachers usually assess (or are expected to assess) these skills.

Data Collection and Analysis

To collect the data (i.e., the assessment instruments), we contacted a group of 30 English language teachers from state high schools in the coffee region of Colombia. They were invited to participate in a professional development course in language assessment. To accomplish this, they agreed to share two assessment instruments with our research group: one assessment for receptive skills and one for productive skills. The teachers shared their instruments by email; then, we deleted all personal and institutional information written in the documents. Finally, we stored the instruments in a digital Google Drive folder, only accessible to us as researchers.

To organize and analyze the content in the instruments, we used two Google Forms (one for listening/reading assessments and one for speaking/writing assessments) which included a description section and an analysis section. For the former, we described the nature of the instruments and their items or tasks: methods, instructions, texts used (for listening and reading), scoring criteria (for speaking and writing), number of tasks or sections, and point allocation.

Regarding the latter, we analyzed qualities of assessments such as validity, authenticity, and reliability. Each one of us scrutinized between thirteen and sixteen instruments. Then, the principal investigator reviewed all the analyses and noted down areas that needed clarification. Comments that needed elaboration were discussed and resolved in the research team. After our analyses, we derived interpretations regarding what was done well and what needed to improve in design.

This content analysis, as commented, was done conceptually and ecologically, i.e., based on emerging trends in the data. We finalized data analysis by grouping content codings, which led us to the major findings we present and discuss below. For example, a major trend in the assessments for listening/reading skills was that many items could be answered without actually listening or reading; speaking and writing assessments were mostly grammar based. These two codes were grouped under one major trend: construct-related issues. As for instruments seeking to assess speaking or writing, we noticed that they did not include clear scoring criteria, or they included construct-irrelevant factors. We labelled these open codings under another major trend: reliability issues.

Findings and Discussion

In line with the research design we explain above, we first present and discuss findings that involve description of the *corpus* of 51 instruments. Following these descriptive findings, we present findings in which we assumed a critical stance towards the *corpus*; this criticality occurred because, as we announce at the beginning of this paper, the examined instruments would lead us to make decisions for educating the participating English teachers through LAL. For both types of findings, we discuss possible reasons for their nature.

Language Skills Addressed and Methods in Corpus of Instruments

From the *corpus*, we identified that most assessment instruments were designed to assess reading skills, with a total of twenty instruments. The next most frequently addressed skill was writing, with a total of thirteen. For listening skills, there were ten instruments and, finally, there were eight for speaking skills.

Regarding assessment methods, for receptive skills, the True-False format was used in ten instruments. Multiple-choice questions and short answer items were included in five instruments. Finally, matching, ordering, and diagrams were used in five instruments. On the other hand, to assess productive skills, six instruments included an analytic rubric and two a holistic rubric. The rest of the instruments —thirteen out of a total of 21— did not include any kind of method to assess performance in speaking or writing: They only included task instructions.

Most instruments targeted reading skills, probably because this is a major component in the English paper of the national standardized test for high schools in Colombia: *Pruebas Saber*. This test

does not include sections for listening or speaking, which may also explain the limited number of assessments for these skills. Additionally, perhaps teachers see that assessing reading is more practical than assessing listening or speaking. For listening, an audio playing device is needed; for speaking, time and other resources are required. For reading, teachers can use readily available texts and design items based on their existing LAL; however, the construction of items for reading and listening was an issue. For instance, the item below was meant to assess listening; however, it can be answered without listening to the source text. We further explain these problems in the analytical section of the findings.

Ins.7, Listening T-FB

B. A person shouldn't walk all day is she is/was in the desert. (True/False)[2]

A learner may answer correctly this true-false item based on logical reasoning and not referring to his/her listening skills: It is true that a person should not walk all day in the desert. Upon analysis of the source text (the video for this listening task), the answer, indeed, was true.

Even though writing was the second most frequently addressed skill in the assessments, many of these instruments (eight out of thirteen) were designed to assess grammar, as the sample below shows; in other words, the teachers shared grammar tests that they labeled as tests of writing skills. Additionally, and perhaps the major issue in productive skills assessment, was the lack of rubrics, as we mentioned earlier. When asked about rubrics, teachers confirmed that they did not have any.

Ins. 36, Writing

Revision: Present simple or present continuous?

Fill in the correct form of the verbs

Look! Tom ___ his bike over there. (ride)

Two (out of eight) speaking assessments included a rubric. However, upon analysis, we noticed that this instrument did not represent crucial areas of speaking. For instance, the sample below is focused on pronunciation, fluency, and visual aid. However, grammar, vocabulary, coherence or other aspects are not present in this rubric; additionally, visual aids are assessed, which are not part of a learner's speaking skills.

Ins. 2, Speaking

Instructions: The students will listen to an audio recorded by the teacher. Then, the students will memmorize the audio to be presented in front of the class, supported material is needed to engage the audience.

Table 5. *Sample Rubric to Assess Speaking*

Pronunciation	The pronunciation is clear, and no mistakes are committed.	2.5
Fluency	The students present no hesitation.	1.5
Visual aids	Visual aids are clear and consistent with the presentation.	1

Source: Author's own elaboration (2021).

As we state in the theoretical and technical considerations, a clear definition of the skills to be assessed in an assessment is a central consideration. In the case of the assessment of writing/speaking skills, teachers need to spend time in the design of instruments for these skills; this may be a reason why the teachers do not have rubrics for these skills: They may have limited time to do it, as was observed in the study by Frodden et al. (2004). Another reason could be that the teachers in our study did not receive training in this area of assessment. Overall, unclear construct definition seems to be an area in which language teachers struggle with regarding the design of assessments, as has been shown in the existing literature (Frodden et al., 2004; Giraldo, 2018b; Levi & Inbar-Lourie, 2019; Villa-Larenas & Brunfaut, 2022). This issue may be exacerbated as the teachers in our study report that they assess other aspects

2 The mistake "is she" is in the original instrument shared by the teacher.

beyond English language skills, as shown in Table 4: discipline, respect, and other construct-irrelevant factors. This problem with construct definition in assessments should inform language testing courses for teachers, because construct definition is, arguably, one of the main pillars of assessment (Bachman & Palmer, 2010; Fulcher, 2010).

Based on these findings, we inferred that the teachers in this study may benefit from a language assessment program in which they learn (or review) how to design proper items for reading and listening, in a way that they can increase their potential to activate these skills. Also, we concluded that teachers needed to be taught about approaches to construct definition for writing and speaking skills, and they needed to problematize construct irrelevance as a major threat to validity. With these skills in mind, we believed that proper assessment design could lead to discussions about qualities of assessment; for example, a robust rubric can increase reliability and validity because the construct of speaking or writing is carefully defined.

Technical and Theoretical Issues in Construct Representation in the Instruments

In this section, we provide a closer, critical analysis of the nature of the 51 instruments we studied. We start by presenting recurring problems that we identified in the design of items for assessing listening and reading skills. Then, we explain the issues we elucidated in the assessments of speaking and writing skills.

Listening and Reading Assessments

Besides, other technical problems we identified in the assessments for listening and reading skills (in addition to questions that can be answered based on prior knowledge) were items that could not be answered and items whose answer was not totally clear. Item 4 below, meant to assess reading, does not have an answer in the text. Item 13 has options that overlap.

Ins. 16, ReadingT-F4

True (T) or False (F)

Item 4: The students write the English answers on the board. T__ F__

The corresponding segment from the source text reads as follows:

For developing the listening skills, the teacher asks us to listen audios to answer questions like this "How old are you? We pay attention to the answers and write them on a piece of paper.

Thus, there is no evidence in the reading to state whether Item 4 is true or false.

Ins. 25, ListeningMCQ13

Item 13: Bludworth is ___ when Pepys finds him.

A. Angry and dirty

B. Angry and hot

C. Hot and dirty

D. Tired and dirty

Supposedly, the answer was C, but options A and D also have the word *dirty;* option B has the word *hot*, so there is overlap among the options: there is no one single, undoubtedly correct answer.

The technical difficulties in the items above, lead to problems with two qualities of language assessment. Validity can be negatively affected because, if a student answers an item without listening or reading, then the resulting mark or score cannot be interpreted as a demonstration of these skills. In addition, reliability is reduced in items that do not have one clearly correct answer: some students may fail because they choose an answer which is partially true (or false), but that is keyed as the opposite —the performance may be providing reliable information about the student's skill but the scoring itself is not. Also, this problem may exacerbate if students, theoretically, chose a different answer if they took the same test again. In synthesis, all these design glitches in the listening or reading assessments may not allow teachers to make accurate inferences about constructs. Based on their design, the assessments did not accurately and wholly address the constructs of listening or reading.

Speaking and Writing Assessments

We found two major issues with the instruments to assess these skills. On the one hand, the rubrics that the teachers used did not fully state what about these skills was to be assessed; a related problem was the lack of rubrics to assess these skills. According to the rubric in Table 5 below, the teacher was meant to assess *knowledge of the content*, *language grammar and vocabulary*, and *voice*. However, these aspects are not clearly defined, which is a key condition for validity in assessment: it is not clear what specific aspects of grammar or vocabulary are assessed.

Table 6. *Sample Rubric for a Speaking Assessment*

Ins. 40, Speaking:

Knowledge of the content	Superior[3]	Alto	Básico	Bajo
1. Student has shown knowledge of the content				
2. The student has prepared for his/her presentation				
Language, grammar and vocabulary	**Superior**	**Alto**	**Básico**	**Bajo**
3. The grammar mistakes did not complicate the comprehension				
4. The language was clear and easy to understand				
5. The vocabulary was appropriate and varied				
Voice	**Superior**	**Alto**	**Básico**	**Bajo**
6. The pronunciation did not interfere with the message				
	Grade:			

Source: Author's own elaboration (2021).

The next instrument is meant to assess writing skills. In this, the only information we could retrieve were the instructions as there was no rubric to assess performance.

Ins. 53, Writing

Make an infographic to adolescents inviting them to carry a healthy life.

Use images, tips, an interesting title or question.

It can be in pairs or individual.

On the other hand, the other issue with the speaking and writing assessments in the *corpus* was the lack of authenticity. Since there was a tendency to address grammar in the instruments, the teachers proposed tasks that bore a limited resemblance to real-life situations and use of the English language. For example, the writing assessment below does not target writing as it is done in real-life situations.

Ins. 2510, Writing

1. You /finish/your lunch/yet/?

_____

2. Your teacher/plan/this class activities/already

_____

3. Claudia/tell/me /a secret/just/

_____

The writing assessment above presents authenticity issues because in real life situations we do not order preset sentences; instead, we write the sentences from scratch according to our goal (e.g., writing

3 These are words in Spanish in the original instrument. They can be roughly translated as follows: Superior: outstanding; Alto: high; Básico: basic; and Bajo: low.

an email, an essay, etc.); in other words, a context is needed for the writing task to be more authentic.

The findings in our study reiterate those in other studies (Frodden et al., 2004; Koh et al., 2018; Villa-Larenas & Brunfaut, 2022): Teachers with limited training in language assessment design instruments that present issues in reliability, validity, and authenticity. This trend may be problematic in practice because language instruments with design and theoretical problems —like the ones we identified in the present study— may not be useful to document or contribute to language learning.

Based on the areas for improvement, we suggested that a language assessment program for these teachers should focus on two major areas. First of all, the teachers may benefit from a design-based course in which they can carefully evaluate and create instruments based on rigorous design guidelines (see Table 1 for examples). This pedagogical implication for teaching language assessment to in-service teachers is supported by studies that have shown how teachers foster their LAL through assessment design (see Arias et al., 2012 and Kremmel et al., 2018 for examples). To complement this technical aspect, the teachers can study qualities of language assessments (e.g., validity, reliability, authenticity) and how they can be used to improve assessments. These two dimensions —the technical and the theoretical— can be used as arguments to explain to teachers how properly designed instruments are useful to achieve educational purposes.

Limitations

One of the limitations in our study is that we analyzed instruments and their design but did not inquire into what purpose they served when used, i.e., we did not ask teachers whether the assessment was for placement, diagnostic, progress, or achievement purposes. However, as we explain in the findings, instruments with design flaws may not contribute to a given assessment purpose because they may provide unclear or limited information about students' skills, whatever the purpose for assessment is. Another limitation, on the logistical side of the study, is that we only studied two assessment instruments per teacher —in no way can two documents represent a teacher's entire approach to language assessment. Other sources of information, such as interviews, can provide more robust information about their existing LAL; this is something we are doing in the larger research study to which this present paper is aligned. Notwithstanding these limitations, our approach to qualitative content analysis was useful to identify areas for improvement in the *skills* (i.e., design) component of LAL, a major need that teachers have reported in the literature.

Conclusions and Recommendations

As we stated at the beginning of this article, the purpose of this research project was to show the analysis of the assessment instruments of the selected group of teachers in order to obtain information that could contribute to the planning of an LAL course for these participants.

To summarize, we found that teachers tend to give more attention to reading assessments than listening and speaking ones. This may respond to the practical advantages of conducting reading assessments due to the few materials required in comparison with the devices needed to assess listening and the logistical issues when assessing speaking in groups with numerous students and little time. Additionally, this tendency may respond to how national policies state the guidelines to assess students' English proficiency in *Pruebas Saber*. Since this exam does not include listening and speaking, many schools and teachers probably decide to exclusively focus on the skills the exam includes.

For assessing reading and listening instruments, True and False and Multiple-Choice questions were used the most. Nonetheless, there were problems in the design of both kinds of items: overlap in the options of multiple-choice questions; items that were not completely false or true; and items that could be solved without reading or listening to the source material because they were based on general knowledge, or they could be guessed. These flaws led to problems with validity, which seems to

be a crucial area for improvement among language teachers in our study and elsewhere reported in the literature.

Regarding writing and speaking assessments, we found several problems; among them, most writing assessments assessed grammar instead of writing skills; the instruments did not include rubrics or if they did, the rubrics did not state what about these skills was to be assessed. Finally, there was a lack of authenticity due to the tendency to assess grammar mainly. Additionally, the lack of rubrics in the instruments for these skills was a major issue we noticed. This lack impacts the validity and reliability of the instruments because the instruments do not give evidence about learners' development of these skills and there may not be consistent decisions when assessing learners' performance. Thus, the instruments' issues in authenticity, validity, and reliability allowed us to make an informed choice: the need to address these qualities of language assessment in the course for these teachers.

Concerning recommendations, the process of this research study justifies the necessity of training teachers in designing assessments and developing their LAL in general. Therefore, an LAL program should address the following contents: how to design items based on design guidelines; how to define and narrow down a construct; approaches to construct definition for writing and speaking skills, within a task-based approach for assessment; and how to address the qualities of language assessment (validity, reliability, authenticity, etc.).

The analysis of language assessment instruments in this study gave us clear information about what teachers do and, more specifically, what they need to improve in language assessment. Hence, we suggest that instrument analysis be done as *one* source to understand and problematize teachers' LAL.

References

Arias, C. I., Maturana, L. M., & Restrepo, M. I. (2012). Evaluación de los aprendizajes en lenguas extranjeras: hacia prácticas justas y democráticas [Assessment in foreign language learning: Towards fair and democratic practices]. *Lenguaje, 40*(1), 99–126. In: https://doi.org/10.25100/lenguaje.v40i1.4945

Alderson, J. (2010). *Assessing reading*. Cambridge University Press. In: https://doi.org/10.1017/CBO9780511732935

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

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

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Brindley, G. (2001). Language assessment and professional development. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara, & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 126–136). Cambridge University Press.

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

Buck, G. (2001). *Assessing listening.* Cambridge University Press. https://doi.org/10.1017/CBO9780511732959

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

Cushing, S. (2010). *Assessing writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

Davidson, F., & Lynch, B. (2001). *Testcraft: A teacher's guide to writing and using language test specifications*. Yale University Press.

Davies, A. (2008). Textbook trends in teaching language testing. *Language Testing*, *25*(3),327–347. https://doi.org/10.1177/0265532208090156

Fulcher, G. (2010). *Practical language testing.* Routledge.

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

Frodden M. C., Restrepo, M. I., & Maturana, L. (2004). Analysis of assessment instruments used in foreign language teaching. *Íkala, Revista de Lenguaje y Cultura, 9*(1), 171-201. https://revistas.udea.edu.co/index.php/ikala/article/view/3146

Giraldo, F. (2018a). 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. (2018b). A diagnostic study on teachers' beliefs and practices in foreign language assessment. *Íkala, Revista de Lenguaje y Cultura, 23*(1), 25-44. https://doi.org/10.17533/udea.ikala.v23n01a04

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

Giraldo, F. (2021). Language assessment literacy and teachers' professional development: A review of the literature. *Profile: Issues in Teachers' Professional Development*, *23*(2), 265–279. https://doi.org/10.15446/profile.v23n2.90533

Green, A. (2014). *Exploring language assessment and testing*. Routledge.

Hasselgreen, A., Carlsen, C., & Helness, H. (2004). *European survey of language testing and assessment needs. Part one: General findings*. http://www.ealta.eu.org/resources.htm

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing, 25*(3), 385–402. https://doi.org/10.1177/0265532208090158

Inbar-Lourie, O. (2012). Language assessment literacy. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1-9). John Wiley & Sons. https://doi.org/10.1002/9781405198431.wbeal0605

Inbar-Lourie, O. (2013, November). *Language assessment literacy: What are the ingredients?* Paper presented at the 4th CBLA SIG Symposium Programme, University of Cyprus.

Koh, K., Burke, L., Luke, A., Gong, W., & Tan, C. (2018). Developing the assessment literacy of teachers in Chinese language classrooms: A focus on assessment task design. *Language Teaching Research, 22*(3), 264–288. https://doi.org/10.1177/1362168816684366

Kremmel, B., Eberharter, K., Holzknecht, F, & Konrad, E. (2018). Fostering language assessment literacy through teacher involvement in high-stakes test development. In D. Xerri & P. Vella Briffa (Eds.), *Teacher involvement in high-stakes language testing* (pp. 173–194). Springer. https://doi.org/10.1007/978-3-319-77177-9_10

Levi, T. & Inbar-Lourie, O. (2019). Assessment literacy or language assessment literacy: Learning from the teachers. *Language Assessment Quarterly, 17*(2), 168-182. https://doi.org/10.1080/15434303.2019.1692347

Luoma, S. (2004). *Assessing speaking*. Cambridge University Press. https://doi.org/10.1017/CBO9780511733017

Malone, M. E. (2017). Training in language assessment. In E. Shohamy, I. G. Or, & S. May (Eds.), Language testing and assessment: Encyclopedia of language and education (3rd ed., pp. 225–240). Springer. https://doi.org/10.1007/978-3-319-02261-1_16

Montee, M., Bach, A., Donovan, A., & Thompson, L. (2013). LCTL teachers' assessment knowledge and practices: An exploratory study. *Journal of the National Council of Less Commonly Taught Languages*, 13, 1–31.

Schreier, M. (2012). *Qualitative content analysis in practice*. Sage.

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing, 30*(3), 403–412. https://doi.org/10.1177/0265532213480338

Villa-Larenas, S., & Brunfaut, T. (2022). But who trains the language teacher educator who trains the language teacher? An empirical investigation of Chilean EFL teacher educators' language assessment literacy. *Language Testing*. Advance online publication. https://doi.org/10.1177/02655322221134218

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