ARTICLE

Exploring the principles of English assessment instruments

Claudio Díaz Larenas [a]
Alan Jara Díaz [b]
Yesenia Rosales Orellana [c]
María José Sanhueza Villalón [d]

Abstract

The instruments language teachers employ to assess student learning are rarely studied and they constitute a significant source of input of how learning and teaching are conceived. The aim of this research is to analyze 205 assessment instruments created by English teachers. This is an exploratory case study, in which the assessment principles of Authenticity, Validity, Fairness, Reliability and Practicality were analyzed within the context of the assessment instruments. The 205 assessment instruments were analyzed by using an analytic rubric, which considered the language assessment principles as criteria. Through the different analyses, it is possible to conclude that traditional assessment was favoured over authentic assessment and four different clusters reveal that language assessment principles manifest in different degrees in each type of instrument. Interestingly, although language learning is mainly about how people try to communicate with others, teachers are still stressing the assessment of grammar and vocabulary knowledge instead of helping students develop the skill of foreign language communication through key authentic assessment, self-assessment and peer-assessment techniques and procedures.

Keywords: Language Assessment. Assessment Instruments. Assessment Principles. Teachers. English.

[a] Universidad de Concepción, Concepción, Chile.
[b] Universidad de Concepción, Concepción, Chile.
[c] Universidad de Concepción, Concepción, Chile.
[d] Universidad de Concepción, Concepción, Chile.

Received: Apr 25 2020
Accepted: Feb 24 2021

1 Introduction

As part of the learning process, teachers design and use assessment instruments to identify learners' English proficiency levels and needs. The design of assessment instruments involves a variety of key aspects, which are sometimes challenging to achieve: context, class size, content, among others. To be able to design an assessment instrument is a professional competence that all teachers should master.

Chilean teacher education guidelines, which set the nationwide standards for all teacher education programs, recognize the ability to design, use and evaluate assessment instruments as one of the several standards all teachers, regardless of the discipline, should master before leaving university preparation; however, research (MARTINIC; VILLALTA, 2015; VERA SAGREDO; POBLETE CORREA; DÍAZ LARENAS, 2018) has shown that Chilean inservice teachers always claim that they do not feel confident when they have to assess their students because assessment involves being able to collect, analyze and report data to students, and not all teachers feel they are prepared to do so.

This study aims to analyze different types of language assessment instruments created by Chilean inservice teachers to examine how the principles of Authenticity, Validity, Fairness, Reliability and Practicality are employed, and classified into describable clusters[1].

2 Literature review

Language assessment always draws teachers' attention and interest because they constantly design instruments, techniques and procedures in an effort to effectively assess what students have learned or provide the kind of feedback they need to enhance their learning. Assessing learning should actually be a highly structured and important activity aligned to certain guidelines and principles that can ensure that assessment really mirrors teaching and can also guide teachers in the design of high quality instruments. Sections 1 to 7 below address some of these key principles.

2.1 Language assessment

Assessment has always been a subject of debate for language experts and teachers. According to Chandio and Jafferi (2015) "assessment is a continuous process which helps both teachers and learners to determine whether the teaching and

[1] This paper is in the context of the research grant FONDECYT 1191021 entitled *Estudio correlacional y propuesta de intervención en evaluación del aprendizaje del inglés: las dimensiones cognitiva, afectiva y social del proceso evaluativo del idioma extranjero.*

learning process is effectively being incorporated" (p. 154). Brown (2004) views assessment as an ongoing process, because "whenever a student responds to a question, offers a comment, or tries out a new word or structure, the teacher subconsciously makes an assessment of the student's performance" (p. 4). Assessing students' progress is one of the vital decisions an educator must take (FRANGELLA; MENDES, 2018) because assessing others involves collecting, analyzing and reporting data for diagnostic, progress, placement or achievement purposes (REYNISDÓTTIR, 2016).

Language assessment has undergone two broad paradigms. Assessment of learning, also known as traditional assessment, refers to practices that involve the use of assessment for administrative purposes (assigning grades, selection, decisions) and looks into what learners can do at the end of the teaching and learning process to rank their achievement levels against a standard. Assessment for learning, also known as authentic assessment, embeds assessment processes throughout the teaching and learning process to constantly adjust teaching and inform learners of how they may improve. In real life, teachers employ both assessment paradigms and regard them as opportunities to gather insights into learner abilities. Regardless of the assessment paradigm, teachers have to design assessment instruments that provide them with the information they need to gain about the students they work with. In designing those instruments, whether for traditional or authentic assessment, teachers should follow a series of principles to create "appropriate assessment" (MCCRAY; BRUNFAUT, 2016).

Brown (2004) stated that these principles need to be applied to all kinds of assessment from tests and quizzes (traditional assessment), to oral presentations and debates (authentic assessment). In fact, it is required that all teachers, regardless of the discipline, should be able to balance these principles in any assessment they design. These five assessment principles are key to assess learners fairly and effectively: Authenticity, Fairness, Practicality, Reliability, and Validity (QIAOCHAN, 2018; TOFFOLI *et al*., 2016).

2.2 Authenticity in assessment

Authenticity in assessment refers to the use of activities and items that reflect real life practices using language as natural as possible, contextualized items, meaningful topics, items provided with thematic information and real-world tasks. The goal of authenticity is that learners can be able to solve real world tasks with their own background knowledge (BURTON, 2011). Using authentic assessment tasks is key to provide students with a meaningful exposure to the language in use, because authentic tasks carry several benefits for learners

such as: rising their motivation, enriching their learning experiences by using real-world English and acquiring language to be used outside the classroom (FREY, SCHMITT; ALLEN, 2012; TOMLINSON; MASUHARA, 2017; VOGT; TSAGARI, 2014).

2.3 Fairness in assessment

Fairness is defined as the application of the same set of rules, standards and criteria in a certain assessment situation to reduce bias in the educator's decision-making. Kunnan (2004) indicated that fairness implies that all learners must complete all assessments in equal conditions considering their background. Assessment is fair when the assessment process is clearly understood by learners and agreed by both assessors and learners and when learners' needs and characteristics are addressed (FULCHER; DAVIDSON, 2007; GREEN, 2014; KUNNAN, 2004).

2.4 Practicality in assessment

According to Coombe (2018), practicality encompasses two main ideas, which complement itself in the practice of this feature. The first one specifies that practicality is the ability of selecting, among certain types of assessments, the best ones, according to the context of the school, it refers "concretely to the teacher or institution's ability to administer the assessment within the constraints of time, space, staffing, resources, government/institutional policies, and candidates or parents' own preferences, among others" (p. 34). The second idea highlights that "all resources available to developers and users of language assessments in the processes of developing, administering, scoring, and using their assessments" (p. 34) must be present in an assessment to achieve practicality. Those resources include human, material and financial assets as well as the time schedule for assessment activities (SCHONLAU; GWEON; WENEMARK *et al.*, 2019). Practicality is a must when creating any assessment. An excellent test may be considered impractical due to a series of issues that have a direct relation with practicality. For instance, time (extension of a test and scoring procedures), price (when we consider the different school contexts), and administrative procedures are just a few of the issues presented when practicality is not considered (SCULLY, 2017).

2.5 Reliability in assessment

Hubley (2007) explains that reliability refers to the consistency of the assessment score. For instance, if an assessment is given to a group of students at a certain

point of time, it will not matter if the same assessment is given to a different group of students. The results should be similar for both groups. Coombe (2018) describes reliability as "consistency of scores or test results" (p. 36). Reliability must be presented if two questions are to be raised: "Are test results dependable and trustworthy? And if a student took the same test the following day would the test results be the same?" (p. 36).

There are some aspects, which may affect reliability; for instance, the format and administrative issues (time, noise, light, seating arrangements). Affective aspects on the part of learners must also be taken into consideration, such as: tiredness, personality and learning styles (COOMBE, 2018).

2.6 Validity in assessment

Validity is related to reliability, as both principles together depend on the assessment and the contents learned by students. According to Hubley (2007), validity measures what the assessment measures. Validity is the feature that confirms whether the assessment has an appropriate complexity level for the students. There are different types of validity in language assessment: content validity, construct validity and face validity. Content validity refers to "how the test assesses the course content and outcomes using formats familiar to the students" (HUBLEY, 2007, p. 22). Construct validity is the "fit between the underlying theories and methodology of language learning and the type of assessment" (HUBLEY, 2007, p. 22), and face validity refers to "the test looks as though it measures what it is supposed to measure" (p. 22).

Validity addresses the procedure of establishing the assessment boundaries and to which extent the assessment evaluates what claims to measure. Hakuta and Jacks (2009) also conceive that validity will show how good an assessment is for a particular situation and will also give meaning to the assessment scores.

In brief, all these five language assessment principles are technical properties of any assessment that indicate the quality and usefulness of the test.

2.7 Public policy for teacher education in Chile

To become a teacher of any subject in Chile, candidates must undergo a process of university preparation that takes between four and five years. University curriculums must be designed and implemented based on pedagogical and disciplinary standards set by the Ministry of Education (CHILE, 2008) and teacher education programs also have to be scrutinized by the National Accreditation

Agency (CHILE, 2019), which also sets standards for quality education. Besides, once graduated, teachers have to be assessed every so often to demonstrate their teaching capacities.

Regardless of the public policy for teacher education, each one of them targets at one specific standard that all preservice and inservice teachers in Chile must achieve, that is, the ability to design, implement and evaluate assessment instruments; in other words, teaching professionals in Chile have to be able to design assessment instruments that meet the technical components of the five assessment principles described above. Therefore, this study examines how these principles are evidenced in the assessment instruments provided by the research participants.

3 Research design

This is an exploratory case study that addresses the design of assessment instruments created by Chilean inservice teachers in terms of the five assessment principles described above. The purpose is to gain a deeper understanding of the design of assessment instruments that were provided by English teachers.

3.1 *Corpus*

The *corpus* is composed by 205 instruments which were provided by English teachers from different educational establishments in the Southern city of Concepción, Chile: ten teachers were from subsidized schools (partially government funded), ten teachers were from public schools (funded completely by the government) and two were university teachers. All these inservice teachers teach English to learners of all ages. They were contacted through email to take part in this research over the second half of 2019, and they voluntarily shared some of the assessment instruments they regularly use with their students. The only specific criterion was that the instrument should have been created by themselves. Table 1 shows the detailed number of assessment instruments provided.

Table 1 - Assessment instruments by type

Types of assessment instrument	Total
Test	124
Numerical rating scales	28
Analytic rubrics	22
Quizzes	9

Continue

Continuation

Types of assessment instrument	Total
Holistic rubrics	6
Test speally designed for SEN students	6
Rating scales	3
Checklists	3
Checklists for self-assessment	1
Checklists for peer-assessment	1
Analytic rubric for self-assessment	1
Test + Rubric	1
Total	**205**

Source: Authors own elaboration (2020)

3.2 Instrument

An analytic rubric was used to analyze each of the 205 assessment instruments collected. The rubric was composed by criteria, indicators and levels of performance. The criteria were the language assessment principles of Authenticity, Fairness, Practicality, Reliability and Validity. Authenticity contained seven indicators: Contextualization, Cognitive level, Content, Interactivity, Language systems/skills, Relevance and Tasks. Fairness showed two indicators: Accessibility and Relationship participant-instrument. Practicality had the indicator of Length. Reliability contained six indicators: Language, Lay-out, Modelling, Scoring, Total score and Specifications. Validity split into five indicators: Instructions, Complexity level, Concordance between items and skills, Correction and Specifications. There were three levels of performance: Satisfactory (3 points), Partially satisfactory (2 points) and Unsatisfactory (1 point).

3.3 Data analysis

Using the analytic rubric, two Chilean expert judges rated each one of the 205 assessment instruments individually. Then three panel sessions were held for the two judges to reach an agreement of their final score for each instrument. The two expert judges were experienced language teacher educators who worked in a Chilean university, training English language preservice teachers. To analyze the judges' agreement, the Cohen's Kappa coefficient was used to determine the strength of the concordance in the different language assessment principles and instruments. Cohen's kappa coefficient measures inter-rater agreement for qualitative items. Cohen's kappa works on a range from -1 to 1. If the results are close to 1, then the concordance between two judges is close. The two judges' agreement levels were between 0.63 and 1.0 in this study.

Then all the data were statistically processed to group the assessment instruments based on their similar characteristics through Cluster analysis. Based on the mean scores for the five language assessment principles, the grouping of instruments yielded four clusters.

4 Findings

From the 205 assessment instruments created by the participants, 12 different types were identified (See Table 1 above). Table 2 below describes the mean scores of the types of assessment instruments under consideration. It is observed that most of the instruments have an adequate score (over 1.50). In addition, the instruments that show the highest scores correspond to: Tests + Rubric (2.10), Rating scales (2.08), Quizzes (2.07), Tests specially designed for students with special needs (2.05), Tests (2.03) and Analytic rubrics (2.03). On the other hand, the types of instruments that receive an unfavourable mean score correspond to: Checklists for self-assessment (1.20), and Checklists for peer-assessment (1.34). With regard to the language assessment principles, it is observed that the highest scores concentrate on the principles of: Authenticity (2.12) and Practicality (2.23). The lowest score focuses on Reliability (1.36).

Concerning Authenticity, the instrument that has the highest score is Checklist (2.59), while the instrument with the lowest score is Checklists for self-assessment and Checklists for peer-assessment (1.64). With respect to Fairness, Tests specially designed for students with special needs have the highest score (2.40), and Checklists for self-assessment and Checklists for peer-assessment have the lowest score (1.00). As for Reliability, the highest score is for Tests + Rubric (2.50) and the lowest score is shared by Rating scales, Numeric rating scales, Checklists, Checklists for self-assessment, Checklists for peer-assessment and Analytic rubrics for self-assessment (1.00). Regarding Validity, the highest score belongs to Quizzes (2.19), and the lowest score belongs to Analytic rubrics for self-assessment (1.43). Finally, in the Practicality principle, the highest score is for Rating scales (2.70) and the lowest score is for Checklists for peer-assessment (1.20).

As for the types of instruments, Rating scales have the highest score in terms of Practicality (2.70) and their lowest score is in Reliability (1.00). The Numerical rating scales have the highest score in Authenticity (1.92) and their lowest score in Reliability (1.00). Regarding Checklists, the highest score is in Authenticity (2.59) and their lowest score is in Reliability (1.00). Checklists for self-assessment have a relatively high score in both Validity and Practicality (1.86), and their lowest score is in both Fairness and Reliability (1.00). Checklists for peer-assessment have a high score in Validity (1.86) and their lowest score is in both Fairness and

Reliability (1.00). Quizzes have their highest score in Practicality (2.33) and their lowest score in Authenticity (1.85). Analytic rubrics have their highest score in Authenticity (2.55) and their lowest score in Reliability (1.36). Analytic rubrics for self-assessment have their highest score in Authenticity (2.64) and their lowest score is in Reliability (1.00). Holistic Rubrics have the highest score in Practicality (2.22) and their lowest score in Reliability (1.42). Tests have a high score in Practicality (2.52) in contrast to Reliability, which has the lowest score (1.57). Tests specially designed for students with special needs have their highest score in Practicality (2.56) and their lowest score in Reliability (1.40). Finally, Tests + rubric have the highest score in Reliability and Practicality (2.50), and the lowest score in Fairness (1.50) (See Table 2).

Table 2 - Mean scores according to assessment instruments

Type of instrument	Authenticity	Fairness	Reliability	Validity	Practicality	Mean Score
1. Rating Scales	2.36	2.25	1	2.11	2.7	2.08
2. Numerical Rating Scales	1.92	1.53	1	1.89	1.88	1.64
3. Checklists	2.59	2	1	1.57	2.23	1.88
4. Checklists for Self-assessment	1.64	1	1	1.86	1.86	1.2
5. Checklists for Peer-assessment	1.64	1	1	1.86	1.2	1.34
6. Quizzes	1.85	1.86	2.11	2.19	2.33	2.07
7. Analytic Rubrics	2.55	1.86	1.36	2.05	2.34	2.03
8. Analytic rubric for Self-assessment	2.64	2	1	1.43	2.2	1.85
9. Holistic Rubrics	2.22	1.88	1.42	2.1	2.42	2
10. Tests	2.01	2.04	1.57	2.02	2.52	2.03
11. Test speally designed for SEN students	1.99	2.4	1.4	1.89	2.56	2.05
12. Tests + Rubric	2	1.5	2.5	2	2.5	2.1
Global mean score	**2.12**	**1.78**	**1.36**	**1.91**	**2.23**	**1.86**

Source: Authors own elaboration (2020)

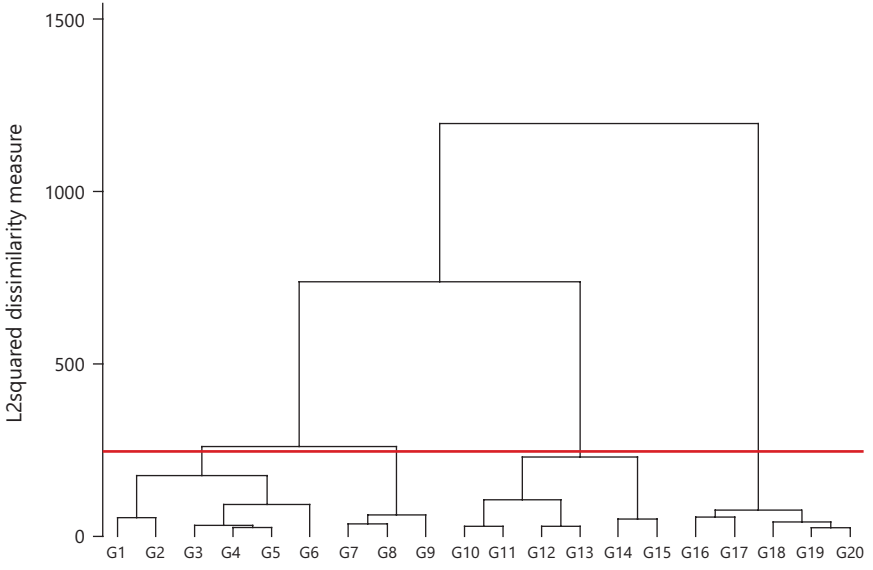
4.1 Cluster analysis

To conduct the cluster analysis, a combination of hierarchical and non-hierarchical methods was used. To identify the ideal number of groups, the hierarchic Ward method was used, based on the calculation of the Euclidean distance to the square

between the standardized information of every evaluation criterion. Moreover, the non-hierarchical method of K-means was used to contrast the grouping obtained from the hierarchical method. The comparison between the results of the hierarchical and non-hierarchical analysis was evaluated through Cramer's V statistics, considering a coefficient > 0.30 to identify an appropriate association (MARTINIC; VILLALTA, 2015). The classification of the instruments was defined on the basis of the evaluation criteria: the language assessment principles analyzed in the literature review.

The dendrogram in Figure 1 illustrates the cluster solution obtained through the application of the hierarchical *Ward* method. A dendrogram is a diagram representing a tree, which shows the hierarchical relationship between clusters. Cluster analysis groups variables and aims to merge the clusters based on their homogeneity. According to the reading of the dendrogram, the structure starts forming twenty groups (G1-G20), and then it keeps diminishing the number of clusters by grouping the assessment instruments according to their characteristics. The final purpose is to reach a point where the number of clusters contains a similar number of instruments to analyze them in a more limited and simpler way. All what is joined together in the dendrogram below the horizontal red line are the clusters. In this case, the dendrogram indicates that four clusters are the most convenient solution.

Figure 1 - Dendrogram for the cluster hierarchical analysis

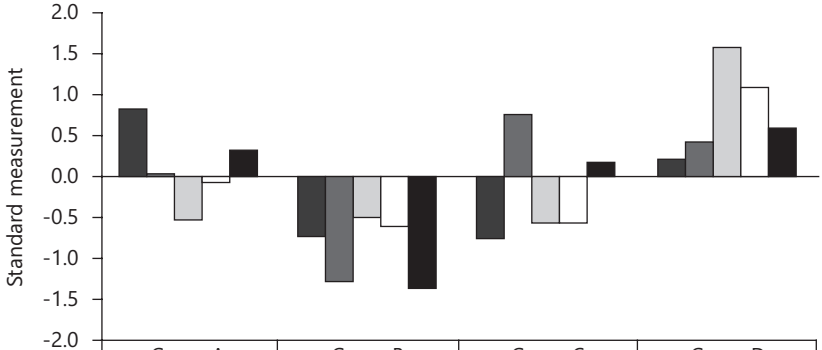


Source: Authors own elaboration (2020)

4.2 Non-hierarchical cluster analysis

Figure 2 below identifies several groups through the non-hierarchical method of K-means, considering the four-group solution suggested from the previous hierarchical cluster analysis. This analysis aims to identify group cases slightly similar, based on their characteristics by using an algorithm that can analyze a great number of cases. The results support the solution of the four proposed clusters beforehand.

Figure 2 - Clusters according to evaluation criteria



	Group A	Group B	Group C	Group D
Authenticity	0.83	-0.75	-0.77	0.22
Fairness	0.02	-1.29	0.75	0.42
Praticality	-0.55	-0.49	-0.58	0.58
Reliability	-0.07	-0.63	-0.59	0.10
Validity	0.32	-1.38	-0.17	0.60

Source: Authors own elaboration (2020)

Figure 2 explains in detail the four clusters formed and their scores in the light of the language assessment principles. Each group is represented in Figure 2 with its scores separately to illustrate the differences. It is easy to identify which groups have the highest scores in each language assessment principle. For instance, Group A has the highest score in Authenticity, Group C has the highest score in Fairness, and Group D has the highest score in Practicality, Reliability and Validity. At the same time, it is easy to identify the lowest scores in each language assessment principle. Group C has the lowest score in Authenticity and Practicality, Group B has the lowest score in Fairness, Reliability and Validity.

Group A corresponds to a set of assessment instruments whose items show relatively high scores in Authenticity (0.83) and Validity (0.32). This means that

the assessment instruments are communicatively contextualized, suggesting that the assessment items are natural and meaningful. The assessment instruments also target high levels of thinking and problem-solving skills. Moreover, these instruments focus on ongoing process-based assessment and present a connection between L2 learning and real-world language use. Regarding Validity, the level of complexity is appropriate for the participants. In addition, these assessment instruments show an agreement between the items created and the skills being measured. Finally, these assessment instruments present varied items with specific outcomes. However, group A shows low scores in Fairness (0.02), Reliability (-0.07) and Practicality (-0.05), which means that these instruments are not very appropriate for the participants' characteristics and nor the length given to the instruments. Regarding Reliability, the cluster suggests that the instruments lack appropriate language, have a poor lay-out, lack modelling and have problems with the subjective interpretation of scores. In brief, these assessment instruments are valid and authentic, but they are that fair, practical and reliable.

Group B entails a set of instruments that show low scores in all language assessment principles: Practicality (-0.49), Reliability (-0.63), Authenticity (-0.75), Fairness (-1.29), and Validity (-1.38). These assessment instruments are poorly contextualized and mainly focused on grammar assessment. The assessment instruments do not show a link between L2 learning and real-world language use. The items and tasks are not likely to be used in the real world. Moreover, they are not designed for participants with special needs. In addition, the length, language and lay-out are not appropriate for the participants. These assessment instruments do not provide modelling and the instructions and scoring systems are not clear. In summary, these assessment instruments are neither practical nor reliable, authentic, fair and valid.

Group C is a set of assessment instruments that show relatively high scores in Validity (0.17) and Fairness (0.75). These scores suggest these assessment instruments are appropriately designed to be completed by participants with special needs. They are appropriate for the participants' characteristics. Moreover, the assessment instruments show clear instructions and appropriate complexity levels. Regarding Validity, these instruments show a variety of items with specific outcomes and a match between the items created and the skill being measured. Finally, this group highlights the variety of items and the clear relation with the skills being measured. Group C portrays low scores in Practicality (-0.58), Reliability (-0.59), and Authenticity (-0.77). These assessment instruments lack contextualization and process-based assessment. Moreover, they target neither higher-level thinking nor problem-solving skills. The items and tasks created do

not show any link between L2 learning and real-world language use. The length of the instruments is not appropriate. Regarding Reliability, the instruments lack appropriate language and clear specifications. Modelling is not present and the instruments are poorly designed. The scoring system is unclear. In brief, these instruments are valid and fair, but they are not necessarily practical, reliable and authentic.

Group D highlights a set of assessment instruments which show high scores in all language assessment principles. The highest score is in Practicality (1.58); this language assessment principle suggests that this group stands out when defining appropriate length for the instruments. The second highest score is Reliability (1.10), which means that modelling and score specifications are present in the instruments. The language used and response specifications are also appropriate. The layout is also outstanding. The third highest score is Validity (0.60). The assessment instruments portray varied items and these match the skills being measured. Language complexity level is besides appropriate for the participants. Then, the fourth highest score is Fairness (0.42), which means that the assessment instruments are well-designed for the participants' characteristics and are also appropriate for learners with special needs. The last highest score is Authenticity (0.22). The assessment instruments of this group are mostly contextualized showing a connection between L2 learning and real-world language use. Additionally, these instruments are communicatively contextualized and focused on process-based assessment. This group aims to assess any of the language skills and systems and targets higher-level thinking and problem-solving skills. Therefore, these assessment instruments are practical, reliable, valid, fair and authentic.

5 Discussion

5.1 About the assessment principles

The 2005 assessment instruments provided by 22 different inservice teachers who teach English in educational establishments from an urban city were then grouped into four clear-cut clusters that show how each one of the five language assessment principles is reflected. Interestingly, there are two extreme clusters in which one of them shows assessment instruments that have high levels of validity, reliability, fairness, authenticity and practicality and the other cluster reveals low levels of each one of the principles. The other two clusters exhibit high levels of validity; therefore, validity seems to be the assessment principle that stands out in all four clusters. This indicates that the 22 teachers who designed the instruments ensured that their assessment procedures had really matched their

teaching because teachers should obviously assess what they teach. These findings differ from what Frodden Armstrong, Restepo Maríjn and Maturana Pararroyo (2004) found out in their study in which the instruments analyzed lacked validity since most of them were composed by expected responses in which there was no variation in the items and they did not match the communicative orientation of the language program the instruments were used for.

Regarding the validity levels found in the current study, it is important to remember that the high levels of validity were found among traditional assessment instruments such as tests and quizzes (133 out of 205). Kane (2010) explained that, even though in the validity principle there it is a high chance of having different perceptions of statements from learners or even between teachers, tests and quizzes tend to be explicit and have simple instructions. As they are short, they tend to have straightforward instructions, which do not leave space to secondary perceptions or misinterpretations (ZIMMARO, 2004).

Authenticity and fairness were the assessment principles that were present in at least three of the three clusters. This indicates that inservice teachers are also concerned about including assessment items that are contextualized and reflect real language situations and employing non-discriminatory assessment practices in which all participants have equal opportunities for being assessed (GIRALDO, 2018). Kane (2010) emphasized that the core of assessment is to treat all learners impartially by taking the same assessment under exact conditions, and their assessment performance should be scored under the same set of rules. Any assessment would be catalogued as fair if the teacher takes into consideration learners with special needs (MONTGOMERY, 2002). If adjustments are needed in both instructions and assessment procedures, the teacher must adapt it to the learners who have any learning problem.

Reliability and practicality were the two assessment principles that scored low in at least four clusters. Moskal and Leydens (2000) add that to reach reliability, well-designed assessments should answer *yes* to the following questions: Are the scoring criteria explicit and straightforward? Are the differences between scoring criteria explicit and clear? And would two different students get the same score for a certain response built in the scoring rubric?

The current Chilean findings differ from what Frodden Armstrong, Restepo Maríjn and Maturana Pararroyo (2004) identified in a group of teachers who suggested that the language assessment principles they relied on the most were practicality and reliability. Both principles are closely related to the design of

traditional assessment instruments. Participants explained that with the aim of designing reliable instruments, they preferred objective items, where there was only one correct answer. They declared that objective items did not raise as many scoring problems when assessing productive skills such as speaking or writing. Therefore, to avoid problems with scoring or subjectivity, they opted to design traditional assessment instruments, which also helped them with time management and resources (PARREIRA; PESTANA; OLIVEIRA, 2018; TUFAIL; JAFFERI, 2015; XI, 2010).

5.2 About the types of assessment instruments

Out of the 205 assessment instruments, 12 different types were identified. The most common type was tests (60%). This result was predictable as in Chile the most used assessment instruments are tests. Coombe (2018) described tests as practical since they help teachers to assess and in most of the cases, grade students' performance and give valuable feedback to learners. Tests are an important part of the Chilean educational assessment policy because of their versatility and easiness when creating and grading them.

As for the language system that is mostly assessed, vocabulary items were present in 170 instruments of a total of 205. This tendency of privileging vocabulary over other language systems (grammar and pronunciation) and skills (listening, reading, writing and speaking) is explained by Kalajahi and Pourshahian (2012), who pinpointed that if teachers opt for a vocabulary teaching approach, learners may gain the different skills in a better and simpler way, and the experience of learning a foreign language will be more effective for students, and they will keep motivated to gain mastery in English.

Additionally, from a total of 205 assessment instruments, 158 instruments assessed the writing skill, 92 instruments assessed reading, 39 instruments assessed the listening skill and 35 instruments assessed speaking. Grammar was also in 148 instruments, and pronunciation in 26 instruments. It is necessary to remember that an assessment may contain not one but several language systems and skills to be assessed.

The mostly used items in the assessment instruments were Fill in the gaps (17%), Matching (13%), Multiple choice and Open-ended questions (12%). These findings are partly explained by what Frodden Armstrong, Restepo Marijn and Maturana Pararroyo (2004) suggested when they pointed out that as teachers experience lack of time, they usually have to look for more objective items which are easy to correct and design.

5.3 About teachers' assessment literacy skill

When exploring the five assessment principles in the 205 analyzed instruments in this study, the research focus is on the skill teachers show to construct quality instruments. This skill is tightly linked to the notion of assessment literacy, what teachers know about assessment. Fulcher (2012, p. 125) defines teachers' language testing and assessment literacy as the knowledge, skills and abilities required to design, develop, maintain or evaluate, large-scale standardized and/or classroom based tests, familiarity with test processes, and awareness of principles and concepts that guide and underpin practice, including ethics and codes of practice.

The importance of identifying the degree of teachers' language assessment literacy is of great significance in teachers' performance since it can enhance or limit student achievement. Muhammad and Bardakcçi (2019) pose that teachers' preparation programs are not well organized to address teachers' needs for classroom assessment. In fact, several studies (MERTLER, 2009; MUHAMMAD; BARDAKCÇI, 2019; YAMTIM; WONGWANICH, 2013) report that most teachers have classroom assessment literacy at a low level.

It is then important to highlight that the examination of assessment principles in a group of 205 instruments is a first window to unfold how literate teachers are when designing assessment instruments. It is true that different assessment paradigms stress certain principles over others. Behaviorism, for example, was very much concerned with Validity and Reliability as two principles that must have been present in all assessment procedures. Authenticity and fairness were probably the focus of attention of more sociocognitive views; however, this study suggests all these principles reveal teachers' assessment literacy levels, particularly in the Chilean context, in which research (VERA SAGREDO; POBLETE CORREA; DÍAZ LARENAS, 2018) has shown teachers tend to assess in a very traditional way and do not view themselves as skilful when they have to assess their students. Learners, on the contrary, highlight that they rarely know how they will be assessed.

6 Conclusion

The most frequently used assessment instruments and the ones that had the highest scores were Tests + rubric, Rating scales, Quizzes, Tests for students with special needs, Tests and Analytic rubrics, while the types of instruments that received the lowest mean scores were Checklist for self-assessment and Checklist for peer-assessment. These findings lead to two concluding remarks: On the one hand, the group of teachers who designed these instruments tend to emphasize

traditional assessment over authentic assessment, that is to say, the use of tests and quizzes to measure mainly knowledge is still widely common among these teachers. Authentic assessment was least frequently used showing that assessment instances in which learners have to show competence and performance are not that common among these teachers. This is worrying because language learning has to do with the ability to communicate and interact with others. Learners may know some grammar, vocabulary and pronunciation, but do not necessarily know how to put them into practice when they communicate. Tests and quizzes only stress the knowledge of the language but not the use of it in communicative situations. It is not surprising either that the analyzed instruments were mainly oriented to the assessment of the writing skills and vocabulary, which were found in 158 instruments and 170 instruments, respectively.

On the other hand, another interesting concluding remark is the fact that self-assessment and peer-assessment, which are key components of authentic assessment, are among the issues that scored the lowest. This reveals again a long standing and traditional view of assessment, in which the only one who assesses is the teacher. Learners and their peers have no assessment role and have nothing to say, according to this view.

Concerning the five language assessment principles, it was also possible to identify the best evaluated instruments independently. In terms of the Practicality principle, the best scored instrument was Rating scales. For Authenticity, the instrument with the highest score was Checklists (2.59). With respect to Reliability, the highest score was for Tests + rubric. In the case of Fairness, the best scored instrument was Tests for students with special needs. Quizzes were the best assessment instruments in terms of validity (2.19). It is then possible to conclude that these principles are reflected in different degrees among the instruments; in other words, some assessment instruments are, for example, more valid and reliable than others, and some are more authentic and practical than others. Since the consideration of these principles is connected with assessment quality, it is then an important challenge to train preservice and inservice teachers into the mastery of these principles to achieve quality in the design of assessment instruments.

It was also possible to organize the assessment instruments based on their similarities to form clusters and give them recognizable characteristics. The cluster analysis created four groups: Group A showed high scores in Authenticity and Validity, but low scores in Fairness, Reliability and Practicality. Group B showed low scores in every language assessment principle. Group C presented

high scores in Validity and Fairness, but low scores in Practicality, Reliability and Authenticity. Finally, Group D achieved high scores in every language assessment principle.

The four clusters in which the assessment instruments were grouped revealed that the group of teachers who designed them expressed different degrees of assessment literacy, which is reflected on the instruments they had to design. For the design of certain assessment instruments, teachers scored high in all five language principles, but for other instruments, teachers scored high in some principles and low in others. This reveals that efforts have to be made to train preservice and inservice teachers into the design of a variety of instruments, which can be useful to collect key data from students' learning and effective enough to provide students with high quality feedback.

This study can become useful evidence for English preservice and inservice teachers, and teacher preparation programs because it is a window to explore teachers' assessment literacy and identify how prepared they are to assess their students effectively considering that teachers always highlight that their assessment practices, techniques and instruments need improvement, help or support. The inclusion of training and practical workshops on assessments would be key to embody a process of self-reflection and improvements in Chilean English teachers' assessment practices. Besides, Chilean teacher evaluation policy requires teachers to show evidence of their ability to design valid, reliable, authentic, practical and fair assessment instruments that are really effective for collecting data of students' learning processes.

The only limitation this study experienced was the fact that it was a challenging task to collect 205 assessment instruments from teachers. It is revealing that they are not willing to share them; teachers seemed themselves apprehensive about making public and available their assessment instruments, which is perhaps a sign that assessment literacy is so important that innovation and change can be tracked and confirmed by examining teachers' assessment practices, techniques and instruments.

Exploração de instrumentos de avaliação de idiomas e seus princípios

Resumo

Os instrumentos que os professores de línguas utilizam para avaliar a aprendizagem dos alunos raramente são estudados e são uma fonte significativa de informações sobre como a aprendizagem e o ensino são concebidos. O objetivo da pesquisa é analisar 205 instrumentos de avaliação fornecidos por professores de inglês. Este é um estudo de caso exploratório, onde os princípios de avaliação da linguagem analisados foram Autenticidade, Validade, Equidade, Confiabilidade e Praticidade. Estes 205 instrumentos de avaliação foram analisados através da utilização de uma rubrica analítica, que considerou os princípios da avaliação linguística como critérios. Por meio das diferentes análises foi possível concluir que a avaliação tradicional é favorecida sobre a avaliação autêntica e quatro agrupamentos diferentes revelam que os princípios de avaliação da linguagem manifestam-se em diferentes graus em cada tipo de instrumento. Curiosamente, embora a aprendizagem de línguas seja principalmente sobre como as pessoas tentam se comunicar com outras, os professores ainda estão enfatizando a avaliação do conhecimento de gramática e vocabulário em vez de ajudar os alunos a desenvolver a habilidade de comunicação em língua estrangeira por meio de técnicas e procedimentos autênticos de avaliação, autoavaliação e coavaliação.

Palavras-chave: *Avaliação. Ferramentas de Avaliação. Princípios de Avaliação Linguística. Professores. Ingleses.*

Explorando los principios de los instrumentos de evaluación del inglés

Resumen

Los instrumentos que el profesorado de idiomas emplea para evaluar el aprendizaje son raramente estudiados, y ellos constituyen una fuente significativa de input sobre la forma como se concibe el aprendizaje y la Enseñanza. El propósito de la investigación es analizar 205 instrumentos de evaluación proporcionados por profesores de inglés. Este es un estudio de casos exploratorio, en el cual los principios de evaluación del idioma analizados fueron Autenticidad, Validez, Equidad, Confiabilidad y Practicabilidad. Estos 205 instrumentos de evaluación fueron analizados mediante el uso de una rúbrica analítica, la cual consideró los principios de evaluación del idioma como criterios. Mediante los diferentes análisis presentados en este estudio, es posible concluir que la evaluación tradicional es preferida sobre la evaluación auténtica; además, cuatro diferentes grupos revelan que los principios de evaluación del idioma se manifiestan en diferentes grados en cada tipo de instrumento de evaluación. Interesantemente, aunque el aprendizaje de la lengua consiste principalmente en cómo las personas se comunican con otros, el profesorado aún enfatiza la evaluación de la gramática y el vocabulario en vez de ayudar al estudiantado a desarrollar la habilidad de comunicación en una lengua extranjera mediante técnicas y procedimientos de evaluación auténtica, auto-evaluación y co-evaluación.

Palabras clave: *Evaluación. Instrumentos de Evaluación. Principios de Evaluación del Idioma. Profesorado. Inglés.*

References

BROWN, H.D. *Language assessment: principles and classroom practices*. New York: Pearson Longman, 2004.

BURTON, K. A framework for determining the authenticity of assessment tasks: applied to an example in law. *Journal of Learning Design*, Brisbane, v. 4, n. 2, p. 20-28, 2011. https://doi.org/10.5204/jld.v4i2.72

CHANDIO, M.; JAFFERI, S. Teaching English as a language not subject by employing formative assessment. *Journal of Education and Educational Development*, Karachi, v. 2, n. 2, p. 151-171, Dec. 2015. https://doi.org/10.22555/joeed.v2i2.444

CHILE. Consejo Nacional de Acreditación. *Criterios y estándares para carreras de pedagogía en Chile*. Santiago, 2019.

CHILE. Ministerio de Educación. Marco para la buena enseñanza. Santiago, 2008.

COOMBE, C. *An A to Z of second language assessment*: How language teachers understand assessment concepts. London: British Council, 2018.

FRANGELLA, R.; MENDES, J. ["What is a good result?" Inquiring the sense of assessment and its curriculum articulations]. *Ensaio: Avaliação e Políticas Públicas em Educação*, Rio de Janeiro, v. 26, n. 99, p. 296-315, Apr.-June 2018. Portuguese. https://doi.org/10.1590/s0104-40362018002600982

FREY, B.; SCHMITT, V. L.; ALLEN, J. P. Defining authentic classroom assessment. *Practical Assessment, Research and Evaluation*, [s. l.], v. 17, n. 2, p. 1-18, Jan. 2012. https://doi.org/10.7275/sxbs-0829 Era et al.

FRODDEN ARMSTRONG, M. C.; RESTREPO MARÍN, M. I.; MATURANA PARARROYO, L. Analysis of assessment instruments used in foreign language teaching. *Íkala, Revista de Lenguaje y Cultura*, v. 9, n. 1, p. 171-208, Dic. 2004. Era Frodden et al.

FULCHER, G. Assessment literacy for the language classroom. *Language Assessment Quarterly*, v. 9, n. 2, p. 113-132, 2012. https://doi.org/10.1080/15434303.2011.642041

FULCHER, G.; DAVIDSON, F. *Language testing and assessment*: an advanced resource book. New York: Routledge Applied Linguistic, 2007.

GIRALDO, F. Language assessment literacy: implications for language teachers. *Profile: Issues in Teachers' Professional Development*, Bogotá, v. 20, n. 1, p. 179-195, Jan.-Jun. 2018. https://doi.org/10.15446/profile.v20n1.62089

GREEN, A. *Exploring language assessment and testing*. Oxon: Routledge, 2014.

HAKUTA, K.; JACKS, L. *Guidelines for the assessment of English language learners*. Princeton: Educational Testing Service, 2009.

HUBLEY, N. Introduction to issues in language assessment and terminology. *In*: COOMBE, C.; FOLSE, K.; HUBLEY, N. *A practical guide to assessing english language learners*. Ann Arbor: University of Michigan, 2007. p. 13-30.

KALAJAHI, S.; POURSHANIAN, B. Vocabulary learning strategies and vocabulary size of ELT students at EMU in Northern Cyprus. *English Language Teaching*, Oxford, v. 5, n. 4, p. 138-149, Apr. 2012. https://doi.org/10.5539/elt.v5n4p138

KANE, M. Validity and fairness. *Language Testing*, London, v. 27, n. 2, p. 177-182, Mar. 2010. https://doi.org/10.1177/0265532209349467

KUNNAN, A. Test fairness. *In*: MILANOVIC, M.; WEIR, C. (eds.). *European language testing in a global context: Proceedings of the ALTE Barcelona Conference*. Cambridge: Cambridge University Press, 2004. p. 27-48.

MARTINIC, S.; VILLALTA, M. [Time management in the classroom and academic performance in full-day schools in Chile]. *Perfiles Educativos*, México, v. 37, n. 147, p. 28-49, 2015. Spanish.

MCCRAY, G.; BRUNFAUT, T. Investigating the construct measured by banked gap-fill items: evidence from eye-tracking. *Language Testing,* London, v. 35, n. 1, p. 51-73, Nov. 2016. https://doi.org/10.1177/0265532216677105

MERTLER, C. Teachers' assessment knowledge and their perceptions of the impact of classroom assessment professional development. *Improving Schools*, [s. l.], v. 12, n. 2, p. 101-113, June 2009. https://doi.org/10.1177/1365480209105575

MONTGOMERY, K. Authentic tasks and rubrics: going beyond traditional assessments in college teaching. *College Teaching*, Washington, DC, v. 50, n. 1, p. 34-40, Jan. 2002. https://doi.org/10.1080/87567550209595870

MOSKAL, B. M.; LEYDENS, J. Scoring rubric development: validity and reliability. *Practical Assessment, Research and Evaluation*, [s. l.], v. 7, n. 10, p. 71-81, Jan. 2000.

MUHAMMAD, F.; BARDAKÇI, M. Iraqi EFL teachers' assessment literacy: perceptions and practices. *Arab World English Journal*, [s. l.], v. 10, n. 2, p. 431-442, 2019. https://doi.org/10.24093/awej/vol10no2.33

PARREIRA, A.; PESTANA, M. H.; OLIVEIRA, P. Assessing educational leadership: a competence-complexity based test. *Ensaio: Avaliação e Políticas Públicas em Educação,* Rio de Janeiro v. 26, n. 100, p. 890-910, Jul.-Sep. 2018. https://doi.org/10.1590/s0104-40362018002601559

QIAOCHAN, L. Application of task-based language assessment in college English writing teaching. In: INTERNATIONAL CONFERENCE ON ECONOMIC DEVELOPMENT AND EDUCATION MANAGEMENT – ICEDEM 2018), 2., 2018, Dailan, China. *Proceedings* […]. *Advances in Social Science, Education and Humanities Research*, v. 290, p. 404-408, Jan. 2018.

REYNISDÓTTIR, B. B. *The efficacy of authentic assessment*: a practical approach to second language testing. Thesis (Bachelor's degree) – University of Iceland, Iceland, 2016.

SCHONLAU, M.; GWEON, H.; WENEMARK, M. Automatic classification of open-ended questions: check-all-that-apply questions. *Social Science Computer Review*, Durham, v. 20, n. 10, p. 1-11, Aug. 2019. https://doi.org/10.1177/0894439319869210

SCULLY, D. Constructing multiple-choice items to measure higher-order thinking. *Practical Assessment, Research and Evaluation*, [s. l.], v. 22, n. 4, p. 1-13, 2017. https://doi.org/10.7275/ca7y-mm27

TOFFOLI, S. F. L. *et al*. [Assessment with construct-response items: validity, reliability, comparability, and fairness]. *Educação e Pesquisa*, São Paulo, v. 42, n. 2, p. 343-358, Apr.-June 2016. https://doi.org/10.1590/S1517-9702201606135887

TOMLINSON, B.; MASUHARA, H. *The complete guide to the theory and practice of materials development for language learning.* New York: Wiley-Blackwell, 2017.

TUFAIL, M.; JAFFERI, S. Teaching English as a Language not subject by employing formative assessment. *Journal of Education and Educational Development*, Karachi, v. 2, n. 2, p. 151-171, Dec. 2015. https://doi.org/10.22555/joeed.v2i2.444

VERA SAGREDO, A.; POBLETE CORREA, S.; DÍAZ LARENAS, C. Percepción de los docentes chilenos sobre sus perspectivas, habilidades y prácticas evaluativas en el aula. *Estudios Pedagógicos*, Valdívia, v. 43, n. 3, p. 361-372, 2018. https://doi.org/10.4067/S0718-07052017000300021

VOGT, K.; TSAGARI, D. Assessment literacy of foreign language teachers: findings of a European study. *Language Assessment Quarterly*, [s. l.], v. 11, n. 4, p. 374-402, Nov. 2014. https://doi.org/10.1080/15434303.2014.960046

XI, X. How do we go about investigating test fairness? *Language Testing,* London, v. 27, n. 2, p. 147-170, Mar. 2010. https://doi.org/10.1177/0265532209349465

YAMTIM, V.; WONGWANICH, S. A study of classroom assessment literacy of primary school teachers. *Procedia - Social and Behavioral Sciences*, [s. l.], v. 116, p. 2998-3004, Feb. 2014. https://doi.org/10.1016/j.sbspro.2014.01.696

ZIMMARO, D. M. *Developing grading rubrics.* Austin: Measurement and Evaluation Center, 2004 [cited 2020 Apr 9]. Available from: http://bsuenglish101.pbworks.com/f/rubricshandout.pdf

Information about the authors

Claudio Díaz Larenas: Doctor in Education. Director of the Master in Innovation of English Language Teaching, Learning and Assessment. Tenure Professor, Universidad de Concepción. Contact: claudiodiaz@udec.cl

https://orcid.org/0000-0003-2394-2378

Alan Jara Díaz: Bachelor of Arts in English Language Teaching. Certified school teacher, Universidad de Concepción. Contact: alajara@udec.cl

https://orcid.org/0000-0002-8667-5222

Yesenia Rosales Orellana: Bachelor of Arts in English Language Teaching. Certified school teacher, Universidad de Concepción. Contact: yrosales@udec.cl

https://orcid.org/0000-0003-1913-9363

María José Sanhueza Villalón: Bachelor of Arts in English Language Teaching. Certified school teacher, Universidad de Concepción. Contact: msanhuezav@udec.cl

https://orcid.org/0000-0003-4680-3207