Alternative Evaluation of a Traditional Oral Skill Assessment Tool in an English Teaching Program

Andrea Lizasoain C. andrea.lizasoain@uach.cl
Amalia Ortiz de Zárate F. aortizdezarate@uach.cl
Universidad Austral de Chile

Language assessment does not have to be a mysterious and remote activity accomplished in isolation from what is done in the classroom.
Richard-Amato, 2010

Abstract

This article reports the findings of an exploratory study that aimed at knowing students' perception of the validity and reliability of tests administered in an English as a Foreign Language university setting, in order to determine if changes needed to be made. Students were given a written assessment tool to apply the contents learnt in an Applied Linguistics course that focused on assessment of a foreign language. This assignment was the research tool to explore students' perception, which yielded both qualitative and quantitative results. It was found that most students were interested in analyzing the validity and reliability of oral examinations, finding that these tests were valid but entirely reliable. This points to the need of changes in the program related to curricular innovation and, more particularly, to different approaches to assessment.

Keywords: English as a foreign language, alternative assessment, oral skills, reliability, validity

Resumen

En este artículo se informan los resultados de un estudio exploratorio cuyo objetivo fue conocer la percepción de los estudiantes acerca de la validez y confiabilidad de las pruebas administradas en la carrera de Pedagogía de Inglés en un contexto universitario. Lo anterior con el fin de determinar cambios necesarios. Los estudiantes recibieron una herramienta de evaluación escrita para aplicar los contenidos aprendidos en un curso de Lingüística Aplicada que se enfocó en la evaluación de una lengua extranjera. Esta actividad constituyó la herramienta de investigación para explorar la percepción de los estudiantes, la que arrojó resultados cualitativos y cuantitativos. Se encontró que la mayoría de los estudiantes se interesó en analizar las pruebas orales, las que calificaron como válidas pero poco confiables. Lo anterior apunta a la necesidad de realizar cambios curriculares y, particularmente, de aplicar enfoques evaluativos diferentes.

Palabras clave: Inglés como lengua extranjera, evaluación alternativa, habilidades orales, validez, confiabilidad

Tests are always a surprise both for teachers and students. Although the latter usually counts on the contents to be evaluated in advance, and teachers (theoretically) measure their knowledge through formats they are accustomed to, students hardly ever know what exactly they are going to be asked or in what order, for instance. In addition, teachers design tests having certain

answers in mind in order to make sure students have acquired certain knowledge; however, many times teachers are taken aback by some of their students' answers, which, without being wrong, are different to what teachers had expected, a reality that Coombe, Folse and Hubley (2007) have also noticed. The latter occurs especially in the case of alternative assessment. Since many skills are assessed at the same time (apart from specific content knowledge) and every student is unique, results are never the same and they are many times surprising for both students and teachers.

Assessment is a key factor in education. For example, it helps us know whether our students have learnt the specific contents of a course or, in the case of students of a foreign language, whether they have acquired the linguistic elements that will allow them to communicate according to different proficiency levels. Now, both the design of assessment tools to measure language learning by teachers and the ability to efficiently take the tests by students are skills (Coombe et al., 2007), and the development of a skill requires training and practice.

Nonetheless, it is well-known that, on the one hand, assessment has not been taken seriously –in fact, "it's seen as a sub-category of language teaching" (McNamara, 2004, p. 763) –and, on the other, teachers are not prepared to design tests that measure what they have to measure and rather improvise according to the assessment instruments they have experienced themselves as students, because there is not much training during their pre-service and in-service years (Figueras, 2005). This can lead to a lack of validity and reliability in the assessment process. Consequently, language learners usually cannot show what they really know (their competence), teachers do not actually know how far their students have gone (their progress), and tests do not accomplish their function (assessing), among other issues.

Today, tests –especially international– are becoming increasingly important to provide evidence of learners' foreign language proficiency and, in fact, their results can decide on their future: Teachers and students depend on their results to be accepted by universities, be hired, receive grants, travel abroad, be awarded prizes, etc. (Brown, 2004; McNamara, 2004; Richard-Amato, 2010; among others). Therefore, test design should be ideal (valid, reliable, practical, etc.), in order to both reflect students' actual competence and help them develop their 'testing skills'.

Since the analysis of tests –and assessment procedures in general– "is the basis for helpful feedback to students, teachers, and administrators" (Coombe et al., 2007, p. 3), a study was conducted based on an alternative evaluation tool that sought to examine the students' perception of the oral assessment carried out during their four years in the language courses of the English Language Teaching Major (*Pedagogía en Comunicación en Lengua Inglesa*) at the Universidad Austral de Chile. The objectives of this task were the following: Make students apply what they had learnt about language assessment in the framework of the Applied Linguistics course; have them experience alternative assessment themselves and weigh the importance of assessment in general; teach them how to evaluate a process; see if assessment in our program complied with

what some authors have called "principles of assessment"; and, at the same time, as they experienced both types of assessment –traditional and alternative–, give them the tools to choose the one that would best suit their future teaching contexts. After all, "assessment coupled with analysis can improve instruction; assessment alone cannot" (Coombe et al., 2007, p. 3).

A theoretical background for the distinction among evaluation, assessment, and testing is presented below, complemented with a contrast between traditional and alternative assessment, the concepts of competence and performance (Chomsky, 1965), the principles of testing (Brown, 2004), and an overview of the speaking skill and its appropriate assessment. The following section includes the description of the present study, preceded by a contextualization. Finally, results are offered together with a discussion and implications.

1 Theoretical Background

1.1 Evaluation, assessment and testing

It is essential to make a distinction among three concepts that sometimes are used as synonyms: evaluation, assessment and testing (Coombe et al., 2007; Figueras, 2005; Nunan, 1992). The term 'evaluation' is the broadest and it refers to a general appreciation of a program or curriculum. Evaluation implies "not only assembling information but interpreting that information –making value judgments (...)" (Nunan, 1992, p. 185). What we seek to know when we evaluate a program or curriculum is whether it has actually met its goal or it is on the right track to meet it.

'Assessment' is part of evaluation in the sense that it has to do with students and their performance (Brindly, 1989, cited in Coombe et al., 2007). It refers to different ways to obtain information about a student's linguistic level, progress and competence; in other words, we assess them to know what they know or can do with what they know. There are many types of assessment aiming at different goals and carried out through different tests, although assessment is not necessarily made through tests exclusively (Ur, 2010).

A test is an assessment tool "used to gather information about students' [linguistic] behaviour" (Coombe et al., 2007, p. xv). This information is usually translated into scores, which are believed "to define the level of knowledge of a testee" (Ur, 2010, p. 33). A test is the realization of a type of assessment: For example, an aptitude test aims at assessing a student's ability to learn a foreign language.

1.1.1 Traditional and alternative assessment

Pen-and-paper tests are usually considered traditional. There is also a kind of assessment known as alternative assessment (Brown, 2004; Richard-Amato, 2010; Ur, 2010), which is thought to be more authentic (Richards & Renandya, 2002). Self-assessment, portfolios, student-designed tests, student-centered assessment, projects, oral presentations, plays, diaries, etc. are considered alternative kinds of assessment. With alternative assessment, not only linguistic

competence is measured, but also other important skills and competencies, such as critical thinking, collaborative work, and autonomy.

1.2 Competence and performance

One of the drawbacks of traditional testing is that it usually assesses 'performance' rather than 'competence'; in other words, conventional tests usually reflect what students can do at a given moment –influenced by the context and their feelings, such as anxiety or fear,– which does not always correlate with what they really know and are capable of understanding and doing.

It was Chomsky who first distinguished between competence and performance in 1965, stating that people's performance in language does not necessarily reflect their underlying knowledge (Cook & Newson, 2007). Brown (2004) supports this idea by stating that "a bad night's rest, illness, an emotional distraction, test anxiety, a memory block, or other student-related reliability factors could affect performance, thereby providing an unreliable measure of actual competence" (p. 117).

Some people "felt that competence should include not only grammatical sectors but also psycholinguistic, sociocultural, and *de facto* sectors" (Richard-Amato, 2010, p. 30). Thus, Canale and Swain (1980) complemented Chomsky's contribution by saying that competence is composed of *grammatical competence*, which "includes the recognition and use of semantic, syntactic, morphological features of the new language" (Richard-Amato, 2010, p. 33); *discourse competence*, which deals with cohesion and coherence of texts (Coombe et al., 2007); *strategic competence*, which deals with how to cope with language difficulties in order to be able to communicate; and *sociocultural competence*, "which stresses the social rules of language use and relies on the social context" (Richard-Amato, 2010, p. 33). For the speaker to be competent in a foreign language, he must be able to master all of these competences.

It is now believed that competence is better assessed through alternative assessment than traditional tests (Richards & Renandya, 2002). However, in order to make alternative assessment a more formal and objective tool to measure students' accomplishments, teachers may use specially designed rubrics. For example, a rubric to assess reading aloud might include the following criteria: delivery, emotion and effectiveness, diction and articulation, intonation, body language, etc. In other words, for assessment to be fair, it must comply with several requirements.

1.3 Principles of assessment

In that regard, Brown (2004) states that for a test to be efficient and effective, it must comply with five principles: validity, reliability, practicality, authenticity, and washback, although only validity and reliability will be discussed here.

Tests should be built to measure the degree in which a content or skill has been mastered, so their design "involves a careful definition of the domain of knowledge, skill, or ability it is targeting" (McNamara, 2004, p. 763). That is the only way a test can be 'valid'; in other words,

the test must match the contents covered in class and the way they have been taught, which is known as 'content validity'. Validity also has to do with the theories and methodologies underlying a given teaching practice: For example, a communicative language learning approach must be matched by communicative language testing (Coombe et al., 2007). This is known as 'construct validity'.

A test must be 'reliable' too, which means it should yield consistent scores at different times (Coombe et al., 2007). Four factors affect reliability: students, raters, the test administration, and the test itself. 'Student-related reliability' has to do with students' different states of mind at different assessment situations; in other words, the same student could do differently on the same test under different circumstances depending on his/her state of mind. 'Rater reliability' is associated with human errors, subjectivity and bias: The same rater could rate different items following different criteria depending on the time of the day, his or her likes or dislikes, etc. "A fundamental concern in the development and use of language tests is to identify potential sources of errors in a given measure of language ability and to minimize the effect of these factors on test reliability" (*ibid.*). 'Test administration reliability' is related to classroom conditions, the time, and possible distractions students could face when taking a test (Brown, 2004); ideally, students should face the same test administration conditions so that results are consistent. Finally, reliability is also related to the way a test has been designed: If it is too long or if students cannot read it because of the letter size or quality of the print, for instance, it will not yield reliable results.

1.4 The speaking skill and how it should be assessed

Speaking is a common and ordinary activity that allows communication among human beings. However, when we think of speaking as a skill that must be assessed in order to determine a foreign language learner's proficiency level, the panorama changes: Speaking is not a common and ordinary activity anymore and turns into a production skill that serves as a means to assess whether the learner is able to understand and produce in the foreign language or not, which is a difficult task, because assessing speaking involves many "kinds of knowing" (Ur, 2010, p. 120): pronunciation, grammar, vocabulary, fluency, and comprehension, which translate into communicative competence.

In the past, there was a tendency to assess mainly grammar and pronunciation accuracy in oral tests –in fact, when we hear someone speak, we immediately focus on how he or she sounds like first (Luoma, 2004). Nevertheless, "the assessment of spoken language has evolved dramatically over the last several decades from tests of oral grammar and pronunciation to tests of genuine communication" (Coombe et al. 2007, p. 112); in other words, fluency has become more important than accuracy: We want effective communicators rather than models of accuracy.

2 The Study

2.1 Context

The curriculum of *Pedagogía en Comunicación en Lengua Inglesa* at the *Universidad Austral de Chile* encompasses seven six-month English language courses –English Language I to VII– in which the four language skills are developed and assessed, including explicit (and implicit) grammar and vocabulary teaching in the light of a communicative teaching approach. Although, in general, teachers enjoy academic freedom, the language and content objectives of these seven subjects are given in advance and they follow a line in terms of teaching and assessment. On the one hand, lessons are divided into skills, grammar points, vocabulary items, etc., and on the other, teachers apply more or less the same kinds and number of tests. Lessons are based on a textbook, which is complemented with handouts and communicative activities (oral presentations, role-plays, interviews, dialogues, etc.). Regarding assessment, in particular, the seven language courses include several (formative) quizzes and two summative tests along the semester: A midterm test and a final test, which try to assess the four language skills separately (if that is possible).

The English subject consists of eight pedagogical hours (45 min.) a week and it is shared by two or three teachers specialized in different skills. Teachers also share the design of tests including the contents covered in the courses, which, at the same time, usually correspond to those covered by a given textbook. Apart from the two compulsory summative tests described above, every teacher can give formative formal and informal tests at any moment.

Regarding oral examinations, these tests are usually included during midterms and final tests and consist in a dialogue between two students or a conversation among four or five students – depending on the language course. From English Language I (first semester) to English Language IV (fourth semester), students being assessed are organized in pairs for the dialogues they must engage in; while from English Language V (fifth semester) to English Language VII (seventh semester), students are assessed in groups of four to five people while having a conversation. During these formal summative tests, students are required to include vocabulary, grammar structures and linguistic and pragmatic functions covered in class. The teachers leading the courses assess students' performance by means of an institutional rubric (see Appendix A: Oral Assessment Rubric) that measures use of general and specific grammar, vocabulary, pronunciation and fluency.

In the case of the dialogues, the pair of students usually come to a small room and sit at table opposite two or three teachers, choose a strip of paper with a question or situation related to the contents covered in class, and have 2 minutes to discuss with one another the manner they are going to develop the conversation. Then they have to engage in a dialogue for 5 minutes. Regarding the group conversations, the group of four to six students come to the same room and find a 200-300 word text related to the contents discussed in the lessons, and have 5 minutes to

read it in silence, without talking about the text with their classmates; after those 5 minutes, they engage in a conversation for 20 to 25 minutes. Students are never interrupted; they are only observed and listened to by the teachers. Interactions are sometimes recorded to double-check teachers' assessment. After taking the test, students receive a graded rubric within 5 to 15 days with comments on their performance.

These tests take 5 to 10 hours to be given by teachers. Students are usually divided into two groups so tests take place in two separate days. Students typically schedule their test time at least one week before, but they normally take the test some minutes later than expected due to unforeseen delays. After teachers give the tests, they gather together to estimate and discuss students' performance no later than two days after having given the tests. When in doubt, teachers resort to the recordings.

The passing mark in the language courses is usually 4.0 (out of 7.0). The average to reach a 4.0 changes with every course: For instance, in English Language I –offered the first semester of the major– 70% of the test must be correct to reach the passing mark; while in English Language VII –offered the penultimate semester– 80% of the test must be correct; so there is a growing demand of fluency and accuracy, and there is a growing difficulty to reach the passing mark.

2.2 Methodology

This classroom research involved 30 subjects who participated in an Applied Linguistics course.1 One of the units of this course dealt with foreign language learners' skills assessment. The 30 subjects were asked to evaluate the assessment system they had gone through during their university years through the following elicitation technique tool: A 1000-word essay referring to the most difficult and easiest tests they had taken, in which they had to examine the testing procedure of two language skills in the light of the five principles of testing by Brown (2004) (See Appendix B: Guidelines for Final Report). Since "language skills" was one the key concepts given in the instructions to prepare the essay, all of the students automatically referred to their English language subjects.

The task had several goals. On the one hand, students would have the opportunity to put into practice the concepts learnt in the unit on assessment, because they had to decide if the tests they had taken had been valid, reliable, practical, authentic, and with washback. On the other hand, when putting these concepts into practice, not only would they evaluate the assessment system of the major, but they would also realize of the difficulties the design of a test comprises, having an impact on the design of their own tests in the future. Nevertheless, the ultimate goal was to count with students' perception of the oral assessment system in our major in order to identify possible flaws and correct them.

[1] This course is offered during the second semester of the last academic year of the major. This allows students to have a perspective of their experience as university students. Besides, since they have had planning and curriculum courses already, they are supposed to have some knowledge regarding this matter.

2.3 Data analysis

The 30 reports written by the students were read and analyzed in detail. Those reports which focused on oral skills assessment were selected for further study, because it was the most chosen skill among students, despite the fact they had been given the liberty to choose two language skills among all the subjects they had taken. The analysis focused on the oral skills assessment tools evaluated by the students in the light of the five principles of testing, but after reading the reports, we decided to focus on two principles only: validity and reliability, because they were the most controversial. We counted how many students thought that the oral summative tests they had taken were valid and reliable, and how many thought they were not. Qualitative and quantitative data was obtained from this analysis.

3 Results

22 out of 30 (73%) students decided to evaluate the oral examinations given as part of the midterm and final tests in their language courses. Regarding validity, 7 out of 22 students (31.8%) considered these examinations were not valid, since, for example,

> some of the topics to be discussed were familiar, but some of them were not and for some of them it was necessary to have background knowledge. [Besides,] Although we had instances to practice oral skills with native speakers, these conversations were not similar to the oral tests; therefore we were not prepared to what we were going to be asked to do[2] (S1)

Or, "we were not prepared to discuss topics among groups, using the language in a close-to-reality situation" (S3). And "apart from that, I think there was a teaching/evaluation mismatch because some of the questions that students had to choose had nothing to do with the topics studied during the units from the courses" (S21).

However, in general, these tests were thought to be valid (68.2% of students), because they "measured what it had been taught" (S6), "students knew beforehand how the test was going to be (…) Furthermore, the tests presented a reasonable challenge for students according to their levels of proficiency (beginner to advanced)" (S17), and "the topics we had to discuss during oral tests had a reasonable challenge for us" (S15). Besides, students knew the oral assessment rubric well before they were assessed, so they were aware of the criteria teachers were going to use, and this rubric is used consistently throughout the seven semesters their language courses last.

In relation to reliability, the picture changes dramatically since 81% of the students thought oral tests results were unreliable due to different reasons. As regards the four different factors related to this principle (rater, student, test, and test administration), Student 3 commented that "I

[2] Students' comments have not been corrected or edited in any way.

was not motivated and creative when speaking, I was anxious and made many interlanguage inaccuracies such as mistakes, errors and false starts". And Student 5 said that "this type of assessment, for me, was like dying; while speaking in front of my teachers, I felt threatened, and I felt that nervousness, like Roberta Flack's song, was 'killing me softly'."

On the subject of rater reliability, Student 8 pointed out that "the day of the oral test was a long working day for professors and sometimes their mood was not the best when taking the last tests," and Student 21 that "neither should I consider them reliable, as teachers almost never recorded what the students were saying and based their evaluation according to what they were listening at the moment."

Most students forgot to refer to test administration reliability, but some of the comments addressed external noises that affected both teachers' and students' perception, such as "test administration was sometimes affected by the noise generated by the teachers who were evaluating us" (S15), and teachers assessed according to what they listened at the moment "despite the weather conditions and noise produced by workers" (S2).

Test reliability itself was not discussed in the essays, so it can be assumed the matters that worried students the most was rater, student and test administration reliability, maybe because these elements had a direct impact on their marks.

4 Discussion and Implications

One of the objectives of this study was to know students' perception of the assessment tools used in our program, to see if we were meeting the principles of assessment suggested by Brown (2004).

As it was mentioned above, among all the assessment tools students had been given during their university years, the most chosen one was the test the English language courses use to assess the speaking skill. That reflects how important oral examinations are to them; after all, these tests decide on their future. It is mostly these tests which determine if they will make it to the next level or not, and therefore, if they will be able to take other courses whose prerequisite is a determined language level. This reality is not only important to them, but also extremely stressful, because it may mean to pay the fees for an extra year at university, among other inconveniences (life expenses, familial problems, etc.).

However, it was not a surprise to see that most students chose to examine this tool, since they always complain about it both through formal –students' Evaluation of Instruction form– and informal means. Therefore, these results came to confirm teachers' assumptions: Students are not happy with the way their oral skills have been assessed and their complaints were now solidly supported, since they had studied how assessment should be carried out.

In general, they considered that oral tests were valid, since contents and assessment matched. Although some students thought otherwise –maybe because they had missed some lessons,– teachers always design oral tests according to the topics covered in class, the testing experience is based on a communicative approach and there is a testing 'ambient' when students are being assessed; in other words, the principle of content validity is fully met: Teachers look for content, construct and face validity (Brown, 2004; Ur, 2010). What is more, this kind of assessment tool "is considered a valid test format because it is generally held that the test result should reflect exactly the level of the type of skill we are interested in" (Qian, 2009, p. 114). Besides, these are last year students in *Pedagogía en Comunicación en Lengua Inglesa*, with an advanced level of English proficiency; in other words, they should be prepared to perform well in more spontaneous forms of speaking (Basturkmen, 2003), that is, in interactive tasks (Brown, 2004; Ur, 2010), namely, group discussions. At this point, it is also important to mention that this kind of assessment not only allows students to show their linguistic competence in the sense of Chomsky's knowledge of language, but also the other four competences Canale and Swain proposed: Students are able to put (old and new) grammar structures into practice (grammatical competence) in a coherent and cohesive way (discourse competence) following the social rules of language that the given context – in terms of content and situation – demands (sociocultural competence), having several opportunities to express themselves in order to communicate their ideas (strategic competence).

However, on the whole, oral examinations results were not perceived as reliable, which coincides with studies that show "that [in general] there is a lack of appropriate, valid, and reliable assessment measures for English language learners" (Valdés & Figueroa, 1996, as cited in López & Bernal, 2009, p. 66). This occurs not only because of the above described factors, but also because assessing speaking is special: "test discourse is not entirely predictable, just as no two conversations are ever exactly the same (…). There is also some variability in the rating process because it involves human raters" (Luoma, 2004, p. 170). However, our major has looked for strategies to tackle these inconveniences. One of them is the use of questions and texts that guide the students towards the use of specific linguistic elements (vocabulary, structures, functions, discourse markers, etc.), thus controlling their performance up to a certain point. Another strategy is the use of a standard rubric across the language courses so as to avoid bias or subjectivity, again, up to a certain point.

Nevertheless, results show that reliability still needs to be approached in order to make assessment even fairer and more consistent. First of all, we need to give students more opportunities to show what they really know, to reflect their real competence in the language. This way, they will also be able to lower their anxiety levels, since they will know that more opportunities to get better results will be available and they will not end up with "the feeling that I had not showed what I truly knew", in the words of a student. Let us not forget that "critical decisions have often been made about second and foreign language students based on their performance on a single language test" (Richard-Amato, 2010, p. 175).

"Assessing so many students simultaneously can lead to inappropriate assessment," as one of the students points out, so professors need to allot more time divided into different days maybe to assess students' oral skills and thus avoid errors due to tiredness and subsequent subjectivity. Errors in assessment can also be avoided if students' performance is recorded. This way, students will also be able to compare the feedback they receive with their actual performance and will count with 'evidence' to discuss their results and marks with their teachers. Nonetheless, there is controversy regarding this topic, because many teachers of the major do record the oral tests, therefore, this argument is not completely reliable. And also, sometimes students ask explicitly not to be recorded because they argue they get more anxious, thus threatening student-related reliability. In consequence, students should be recorded not only during their oral tests, but also in other informal formative tasks so as to get them to be acquainted and more comfortable with the assessment procedures of the major.

It would also be a good idea to find an appropriate place to assess students that is free of noise and other distractions that might affect students' and teachers' performance, thus complying with the principles of test administration reliability. Unfortunately, one of the weaknesses of our major is the lack of suitable facilities to assess our students. However, here it is important to point out that there are certain nuances regarding students' comments. Firstly, students referred to noises made by teachers during their assessment processes; most of the times those noises are not actual interruptions or it does not mean that teachers speak over them. They are rather noises produced by teachers writing on a note pad. Sometimes students feel stressed when teachers stop looking at them and look down to start taking notes, because they think teachers are only writing down the errors they make. Secondly, students also referred to weather conditions arguing that it is detrimental to their performance, affecting the reliability of the test. Nonetheless, it is quite difficult to control the weather in a city like Valdivia, where it rains regularly (10 months out of 12). What are teachers supposed to do if it is raining cats and dogs or else there is a man mowing the lawn? The problem here lies in the facilities, because there should be acoustically isolated spaces where both teachers and students feel at ease in such a stressing situation.

In any case, teachers in the major not only need to improve the design and implementation of oral tests, but they also need to work on the students' skills to take them. Strategies need to be designed to lower students' anxiety during oral examinations and general situations that involve speaking. At the same time, "many EFL teachers find themselves involved in language testing from the very beginning of their careers" (Wharton, 1998, p. 127); therefore, we also need to train our students so they can design fair tests in the future. After all, "the test result is largely determined by the skill of the examiner" (Qian, 2009, p. 115). In this regard, Figueras (2005) states that

> there is no magic bullet and the only way to overcome current poverty of practice is teacher training. Teachers have very little pre-service or in-service training in testing and assessment. And training in the design of testing and assessment programmes, training to

> understand when tests and exams are valid and reliable, and when they are not, training to develop useful classroom assessment and feedback materials, is necessary if we want to bring about better tests and more responsible test use, as well as better teaching related to real life communication needs. (p. 53)

In relation to this author's proposal, another objective of this study was met, since not only did we make our students apply what they had learnt about language assessment in the framework of the Applied Linguistics course, but we also had them undergo alternative assessment themselves and weigh the importance of assessment in general. At the same time, as they experienced both types of assessment – traditional and alternative –, we gave them the tools to choose the method that would best suit their future teaching contexts.

In this light, after this exploratory study, we came to the conclusion that, although our institutional oral skills assessment tool is effective since it is authentic, valid and partially reliable, a further change that was not pointed out by students needs to be made. This change has to do with the rubric we use to assess students' performance during examinations, which focuses mostly on accuracy rather than fluency: Three criteria address accuracy (grammar, vocabulary, and pronunciation) and only one criterion deals with fluency (speed), meaning that it mainly addresses linguistic competence, thus leaving other very important elements of communication aside. In other words, we need to include criteria to assess our students' strategic, discourse, and sociocultural competences in our rubric, not only because they are essential to communicate effectively, but also because these elements are also covered in class. However, it is important to bear in mind that we are training future teachers of English in a foreign setting – which is obviously a disadvantage when learning a language that is not spoken in that setting. Hence, they need to make a greater effort to become accurate when it comes to communicating, since probably the only interaction their future students will have will be with them.

Despite the efforts to improve our way to assess oral skills, we are aware that there is a long way to go. However, we want to highlight the usefulness and fruitfulness that alternative assessment has to offer. Students underwent alternative assessment – and, in consequence, they could contrast this method with traditional testing – and were able to reflect upon their own learning experience and the way they could change assessment procedures in their future careers, so as to get their own students to turn assessment instances into an educational – rather than stressing – practice. Thus, the test becomes a learning experience more than the mere result of that experience, building knowledge instead of overthrowing willingness to learn.

References

Brown, D. (2004). *Language Assessment: Principles and Classroom Practices*. USA: Longman.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* Massachusetts: The MIT Press.

Cook, V.J. & Newson, M. (2007). *Chomsky's Universal Grammar*: *An introduction* (3rd ed.). USA: Blackwell Publishing.

Coombe, C., Folse, K., & Hubley, N. (2007). *A Practical Guide to Assessing English Language Learners*. Michigan: University of Michigan.

Figueras, N. (2005). Testing, Testing, Everywhere, and not a While to Think. *ELT Journal 59*(1), 47-54. doi: 10.1093/elt/cci006

Basturkmen, H. (2003). Beyond the Individual Speaker in New Zealand. In Coombe, C. & Hubley, N. *Assessment Practices* (pp. 91-101). Virginia: TESOL.

López, A. & Bernal, R. (2009). Language Testing in Colombia: A Call for More Teacher Education and Teacher Training in Language Assessment. *Profile 11*(2), 55-70. Retrieved from http://www.revistas.unal.edu.co/index.php/profile

Luoma, S. (2004). *Assessing Speaking.* Cambridge: CUP.

McNamara, T. (2004). Language Testing. In Davies, A. & Elder, C. *The Handbook of Applied Linguistics* (pp. 763-783). Oxford: Blackwell.

Nunan, D. (1992). Program Evaluation. *Research Methods in Language Learning* (pp. 184-210). New York: CUP.

Qian, D. (2006). Comparing Direct and Semi-Direct Modes for Speaking Assessment: Affective Effects on Test Takers. *Language Assessment Quarterly: An International Journal 6(2),* 113-125. doi: 10.1080/15434300902800059

Richard-Amato, P. (2010). Language Assessment and Standards. *Making it Happen. From Interactive to Participatory Language Teaching: Evolving Theory and Practice* (pp. 174-207) (4th ed.). New York: Pearson Education.

Richards, J., & Renandya, W. (2002). *Methodology in Language Teaching: An Anthology of Current Practice*. Cambridge: Cambridge University Press.

Ur, P. (2010). *A Course in Language Teaching: Practice and Theory* (10th ed.). Cambridge: CUP.

Wharton, S. (1998). Teaching Language Testing on a Pre-Service TEFL Course. *ELT Journal* 52(2), 127-132. doi: 10.1093/elt/52.2.127

Appendix A: Oral Assessment Rubric

Oral Assessment Rubric

Language Proficiency in Speaking and Understanding English

I. II. GRAMMAR/ NEW GRAMMAR

5 Precise use of correct structures; few (if any) noticeable errors of grammar or word-order./ Use of many new structures.
4 Consistent use of correct structures; only occasional grammatical or word-order errors which do not, however, obscure meaning (e.g., "I am needing more English", "He gave to me the letter")./ Use of some new structures.
3 Inconsistent use of correct structures; meaning occasionally obscured by grammatical and/or word-order errors./ No use of new structures.
2 Minimal use of correct structures; grammar and word-order unsatisfactory. Frequently needs to rephrase constructions and/or restricts him/herself to basic structural patterns (e.g., uses the simple present tense where the past or future should be used).
1 Grammar and word-order errors make comprehension difficult.
0 Grammatical and word-order errors make speech virtually unintelligible.

III. PRONUNCIATION: (including word accent and sentence pitch)

5 Pronunciation does not interfere with comprehension; speaks with few (if any) traces of "foreign accent". Intonation varied for effect.
4 Pronunciation seldom interferes with comprehension, although one is always conscious of a definite "accent".
3 Pronunciation and poor intonation occasionally interfere with comprehension; "foreign accent" necessitates concentrated listening and leads to occasional misunderstanding. Words and sentences must sometimes be repeated.
2 Pronunciation interferes with comprehension; many serious errors in pronunciation (e.g., "still" sounds like *steel*, "laws" sounds like *loss*), word accent/intonation (words are frequently accented on the wrong syllable), and sentence pitch (statements have the "melody" of questions, etc.). Frequent repetitions are necessary.
1 Pronunciation dramatically interferes with comprehension; very hard to understand due to difficulties with pronunciation, accent/intonation, and pitch.
0 Pronunciation is virtually unintelligible.

IV. VOCABULARY:

5 Accurate and creative use of vocabulary; use of vocabulary and "idioms" is virtually that of a native speaker of English.
4 Rarely has trouble expressing him/herself with appropriate and varied vocabulary and "idioms".
3 Sometimes uses incorrect terms and/or round-about language because of inadequate vocabulary.
2 Limited and/or frequent incorrect use of vocabulary.
1 Misuse of words and very limited vocabulary make comprehension difficult.
0 Vocabulary is inadequate for even the simplest conversation.

V. GENERAL SPEED OF SPEECH AND SENTENCE LENGTH:

5 Speech speed and sentence length are those of a native speaker; speed is varied for effect.
4 Use of appropriate speed; although speed seems to be slightly affected by language difficulties.
3 Use of irregular speed, with both speed of speech and sentence length affected by language difficulties and limitations or by native language habits.
2 Both speed of speech and sentence length are strongly affected by language difficulties and limitations or by native language habits.
1 Speed of speech and sentence length are so far from "normal" as to impede conversation.
0 Speech is so halting and fragmentary, and/or affected by native language habits, as to make conversation nearly impossible.

NAME:	**1**	**2**	**3**	**4**	**5**
I. GRAMMAR	4	8	12	16	20
II. NEW GRAMMAR	4	8	12	16	20
III. PRONUNCIATION	4	8	12	16	20
IV. VOCABULARY	4	8	12	16	20
V. GENERAL SPEED OF SPEECH AND SENTENCE LENGTH	4	8	12	16	20
	20	**40**	**60**	**80**	**100**

COMMENTS:
TOTAL RATING: _____ (100 possible points)

Appendix B: Guidelines for Final Report

Guidelines for Final Report (25% of the course)
Due: November 29th, 2010
Directions:

1. Reflect on all these university years focusing on the **types of assessment** you have experienced.
2. Consider the **types of tests** you have had and classify them into difficult and easy. Explain why some tests have been difficult or easy for you.
3. Choose two of the language skills (reading, writing, listening, and speaking) and analyse how they have been assessed in general terms in the light of the five **principles of testing.** Include the answers to these questions: Do you feel assessment has been fair? What would you keep the same? What would you change? Why?
4. Finally, write a 2-3 page essay with your reflections. Make sure to include the terms below.
   - Respect the school format to hand in written reports (Stylistics). Omit the abstract.
   - Use the APA system to quote and cite.
   - Peer-edit your text before handing it in. Texts full of mistakes will not be read.
   - You will be evaluated with the school rubric to assess writing.

Terms to be included:

- ✓ Competence & performance
- ✓ Inaccuracies (you can mention some)
- ✓ Assessment & evaluation
- ✓ Grammar
- ✓ Correction & feedback
- ✓ Proficiency levels (you don't have to mention them all, but refer to the idea)
- ✓ Error Analysis
- ✓ Teaching/evaluation mismatch
- ✓ Elicitation
- ✓ Principles of testing (refer to all of them)

Please, make the writing a pleasant experience. Take time to reflect on the topic and to write your essays. Remember I will read them and reading should always be a pleasure.