Íkala, revista de lenguaje y cultura
ISSN: 0123-3432
revistaikala@udea.edu.co
Universidad de Antioquia
Colombia

Frodden Armstrong, María Cristina; Restrepo Marín, María Isabel; Maturana Patarroyo, Liliana
Analysis of Assessment Instruments Used in Foreign Language Teaching
Íkala, revista de lenguaje y cultura, vol. 9, núm. 15, enero-diciembre, 2004, pp. 171-201
Universidad de Antioquia
Medellín, Colombia

SIN TÍTULO, 2003
Acrílico sobre papel, 95 x 70 cm.

Analysis of Assessment Instruments Used in Foreign Language Teaching*[1]

Mg. María Cristina Frodden Armstrong**
Est. María Isabel Restrepo Marín***
Espec. Liliana Maturana Patarroyo****

This article presents partial results of a research project on foreign language teachers' discourse and practices with respect to assessment, the aim of which is to improve teachers' assessment practices. The study, conducted in two Colombian universities, has various components: analysis of documents**,** interviews with teachers and students, and workshops with participating teachers in order to qualify them and agree on an improved assessment system. In this report we discuss the analysis made of tests, grids, registers, and forms and other kinds of instruments that teachers use to assess their students. For the analysis of instruments we used an inductive-deductive procedure whereby categories emerging from a first analysis of instruments were then refined by comparing them to those proposed by Bachman and Palmer (1996) and other authors. In general, teachers seem to prefer "hard" over "soft" types of assessment. Moreover, the qualities of assessment on which they seem to rely the most are practicality and reliability; and the ones least taken into consideration are authenticity and interactivity.

Keywords: assessment, qualities of tests, foreign language testing, language competence

Este artículo presenta resultados parciales de un proyecto de investigación acerca del discurso y las prácticas evaluativas de los profesores de lenguas extranjeras cuyo objetivo es mejorar sus prácticas evaluativas. El estudio, llevado a cabo en dos universidades colombianas, tiene varios componentes: análisis de documentos, entrevistas con profesores y estudiantes y talleres con los profesores participantes con el fin de cualificarlos y acordar un mejor sistema de evaluación. En este informe discutimos el análisis que se hizo de los exámenes, parrillas, registros, formatos y otros tipos de instrumentos que usan los profesores para evaluar a sus estudiantes. Para el análisis de los instrumentos utilizamos un método inductivo-deductivo en el cual las categorías emergentes de un primer análisis de los instrumentos se refinaron comparándolas con las propuestas por Bachman y Palmer (1996) y otros autores. En general, los profesores tienden a preferir tipos de evaluación "hard" más que "soft". Además, las cualidades de la evaluación en las que parecen confiar más son la viabilidad y la fiabilidad, y las que menos tienen en cuenta son la autenticidad y la interactividad.

---

* Recibido: 15-06-04 /Aceptado: 03-08-04

1 This research receives financial support from Vicerrectoría de Investigaciones de la Universidad de Antioquia, Universidad Nacional de Colombia, Medellín and Fundación Universitaria Luis Amigó.

Palabras clave: evaluación del aprendizaje, cualidades de las pruebas, pruebas en lenguas extranjeras, competencia lingüística.

Cet article présente les résultats partiels d'un projet de recherche concernant le discours et les pratiques évaluatives des professeurs de langues étrangères dont l'objectif est d'améliorer leurs pratiques évaluatives. L'étude, menée dans deux universités colombiennes, se compose de divers aspects: analyse de documents, entretiens avec des professeurs et des étudiants, et ateliers avec les professeurs participants dans le but de les former et de mettre au point un meilleur système d'évaluation. Dans cet article, on discutera l'analyse faite d'examens, de grilles d'évaluation, de registres, de formats et d'autres types d'instruments utilisés par les professeurs pour évaluer leurs étudiants. Pour l'analyse de ces instruments, on utilise une méthode inductive-déductive qui permet l'amélioration des catégories émergentes d'une première analyse des instruments, celles-ci ont été améliorées en les comparant aux propositions de Bachman, Palmer (1996), ainsi que d'autres auteurs. En général, les professeurs tendent à préférer des évaluations de type "hard" à celles de type "soft". De plus, les qualités d'évaluation auxquelles ils semblent tenir le plus sont la viabilité et la fiabilité et le moins l'authenticité et l'interactivité.

Mots clés: évaluation d'apprentissage, qualités des épreuves, épreuves en langues étrangères, compétence linguistique.

Although evaluation is something that is part of our daily life, it is one of education's most complex issues when assessing students' performance. At the university level, the purposes of assessment as well as its criteria are as many and dissimilar as are teachers (Salinas, n.d.). While addressing the lack of common ground, which may be responsible for the promotion of students to higher levels without empirical justification, a study is being conducted in two language programs attached to two public universities in a large Colombian city. The aim of this study is to identify and describe the assessment discourse and practices of French and English teachers to improve their assessment of student performance.

The study has various components: analysis of documents such as course programs and assessment instruments; interviews with teachers and students; and workshops with participating teachers in order to qualify them and agree on an improved assessment system. In this article we report partial findings based on the first component in an attempt to answer the following research questions: What are the instruments used by participating teachers in order to assess their students? What are the characteristics of these instruments? A follow-up article shall report comprehensive findings of this research.

First, we present a brief description of the context of the study. Second, we explain the research procedures followed. Third, we describe the assessment instruments provided to the researchers by the teachers. Fourth, we analyze "hard" assessment instruments used by participants to assess students' achievement in language courses; and finally, we present an analysis of "soft" assessment instruments.

1. Context

Both programs subscribe to an internationalization strategy whose main purpose is to train students in the accurate and fluent use of a foreign language in order to participate in the global community. These courses are not part of the students' study programs; however, at one of the universities, students are required to have passed four levels of a foreign language in order to graduate.

At the other university, only a minority of students, those with a high GPA, may register; however, all students must pass a reading comprehension exam at the end of their study program. Courses are free, but students who fail pay for the courses when they repeat them. Thus, the consequences of assessment on the students' future educational career may be considerable. The student who does not have the money to pay may stop studying a foreign language, or may have difficulty complying with graduation requirements.

Both programs advocate the use of qualitative evaluation and students are not given a final grade but a pass or fail. Although innovative and alternative assessment practices such as portfolios, projects, and self- and peer-assessment are presently being promoted in both programs, more traditional instruments like quizzes, exams, and drills (*talleres*) are still very much used by teachers, who are the testers, since they are the ones who design, administer, and score the tests. Standardized tests or examinations such as the Test of English as a Foreign Language (TOEFL), Michigan English Language Institute College English Test (MELICET), the University of Cambridge English examinations, *Diplôme d'Études en Langue Française* (DELF) exams, *Test d'Accès au DALF* (TAD) or *Diplôme Approfondi de Langue Française* (DALF) exams are not used to assess students' performance.

Teachers in these programs are all hired by the hour (*profesor cátedra*) and work in different settings to make a living: some teach in two institutions; some in four. Most have four or five groups of students, some even seven or more**;** so the number of students a teacher has to attend ranges between 60 and 180. As stated by one of the participants:

> Quote 1: Colombian teachers, for example the case of the teacher hired by the hour, have to be on the run so much and have to go from one institution to another, and have to make such an effort to earn a decent salary to pay the bills, that they do not have the spirit to come home at night and write. (Interview with Teacher 1)[2]

2 Interviews were conducted in Spanish and translated by the authors. For the original version in Spanish see Annex 1.

Most are *licenciados* — have studied four years at the university to become teachers —are between thirty and forty years old, and have taught basic and intermediate courses to adults. Half of them have attended seminars and/or conferences on assessment.

In the following section**,** we explain the procedures we followed to describe and analyze the instruments these teachers used to assess their students.

2. Research procedures

In order to understand teachers' practices, situations and beliefs regarding assessment, and to devise common assessment guidelines, we adopted a collaborative action research approach. This type of collective enquiry contributes to teachers' professional development and to the improvement of educational practices in the contexts where it is conducted (Altrichter, et al., 1993; Kemmis and McTaggart, 1988; Burns, 1999). Twelve English teachers and five French teachers (out of a total of fifty-one teachers) volunteered and committed themselves to participate in this research by choosing the following options offered: provide the researchers with the instruments used to assess their students (tests, grids, registers, formats, and other kinds of assessment instruments), respond to some interviews, and participate in several workshops to discuss assessment practices. After choosing one or more of these options, participants signed a confidentiality agreement.

First of all, we studied program guidelines and course programs, and asked participants to answer a questionnaire containing personal and professional information. Then we collected tests, grids, registers, formats, and other kinds of instruments that teachers use to assess their students. Teachers provided 106 assessment instruments for the analysis of which we used an inductive-deductive procedure. After an extensive analysis of the instruments where we grouped them per teacher and by type (exams, quizzes, drills, portfolios, papers, etc.), categories emerged to describe test tasks. These categories were refined after we compared and contrasted them with those proposed by Bachman and Palmer (1996), Bustos (1997), the Common European Framework of Reference for Languages (Instituto Cervantes, 2002), and the Multilingual Glossary of

Language Testing Terms (ALTE Members, 1998). With these new refined categories, we devised a chart to analyze instruments and tasks (see Annex 2), taking into account: type of assessment, scoring, rubric, characteristics of the input, characteristics of the student's expected response, and topic.

In the following section we describe the assessment instruments provided by the teachers who participated in this study.

3. Description of assessment instruments

We shall first describe "hard" and then "soft" assessment instruments. "Hard" refers to a traditional way and purpose of assessing which emphasizes objectivity, precision, reliability, and focuses on product rather than on process. It uses mainly quantitative data provided by instruments such as exams and tests (Carroll, 1993). "Soft" assessment, on the other hand, deals with a naturalistic, alternative way and purpose of assessing (Carroll, 1993). Bustos (1997) describes this new tendency as a type of assessment that, by being continuous, takes into account the student's learning process in order to promote it. It is flexible and transparent, and has an assessment characteristic which is intersubjective.

The majority of assessment instruments provided by participant teachers belonged to "hard" assessment, i.e., teachers provided thirty-nine quizzes, twenty exams, and thirteen drills for five different levels.

Quizzes were generally used to assess only one aspect of language (grammar), one receptive skill (reading or listening), or one unit of a textbook or program. They tended to be objective and were usually one page long. Exams assessed different aspects of the language and different skills, and students' achievement on several units of a textbook or program (see Annex 3 for a sample provided by one of the teachers). Unlike quizzes, they assessed a productive skill, e.g., writing; however, they included more items than prompts—"input in the form of a directive, the purpose of which is to elicit an extended production response" (Bachman and Palmer, 1996: 52), which are more amenable to subjective

scoring. They were longer than quizzes—about two or more pages—and were meant to be taken in one complete class session, usually two hours. This is how a teacher described an exam:

> Quote 2: I try for each exam to contain different assessment strategies. So I try to have a matching section, a completion section, another section on writing, another on listening comprehension, or a dictation, or True or False. In other words, I like it to be as dynamic as possible. (Interview with Teacher 1)

For the purposes of this study, a drill was a set of exercises used to reinforce one specific aspect of the language, in or outside class. Drills usually contained objective items that were not contextualized, and included few or no sections on productive skills. Quizzes and drills were part of continuous assessment (*seguimiento*) whereas exams, which are part of summative assessment, were sometimes administered in the middle of the semester, and always at the end. In general, there was a tendency to assess grammar and vocabulary in an objective way.

Regarding the rubrics of assessment instruments, one-third of them stated the institution and one-fourth stated the level to which they correspond. The test's author, the test's name, and the scoring method were rarely written on the tests. Most of them included instructions for each task or section, but very few included general instructions. For example, none of them explained the time allotted to each task, section, or to the whole exam.

The input channel students were expected to process and respond to tended to be visual, usually a written text in the target language. In English tests, the input was sometimes graphic, audio, and audiovisual.

Very rarely was a prompt provided for the type of input. Input was usually in the form of items. Ranking them regarding frequency, the most popular items were: word completion, open question, multiple choice, transformation, and sentence completion. Low in the scale of occurrence were closed questions, correction of errors, translation, matching, dictation, true/false, and classification of words.

Regarding the language of the input, it was characterized by short sentences of ten words or less of the type Subject+Verb+(Object(s) or Complement), some Adverbial+Subject+Verb+(Object(s) or Complement) structures, and by restricted vocabulary. In reading texts we found longer sentences, but usually less than two lines, with broad vocabulary. The type of texts used for reading comprehension in English was usually descriptive and sometimes expository or narrative. In French it was mainly expository. The function of the language most widely used was ideational and manipulative (Bachman and Palmer, 1996), the latter found mainly in the instructions, followed by the imaginative. The topics were usually personal. The language used in the input did not vary regarding sociolinguistic characteristics: it was a standard variety, neutral register, with few or no cultural references, and no figurative language.

The characteristics of the expected response paralleled those of the input. The only difference was in the types of text students were asked to produce. While in English all types of texts were equally required, students were not required to write argumentative texts in French.

So far, we have described some of the teachers' preferred ways of assessing students, the type of tasks, the language of the input, and the expected response of "hard" assessment instruments. "Soft" assessment instruments will be described next.

In the documents provided by the participating teachers, there were few instances of "soft" assessment. "Soft" assessment is related to procedures that are more qualitative in nature and that reveal a formative and process-oriented type of assessment, such as portfolios, interviews, self- and peer-assessment, role plays, and papers. These alternatives in assessment (Brown and Hudson, 1998) aim at making the assessment of the learning process more democratic and fairer as decisions are made on the basis of different sources of information and not only on tests.

Some English teachers provided written instructions for in-class tasks, i.e., communicative activities carried out by the students with a specified objective, procedure, and outcome. These activities are as close as possible to the communicative tasks that the students will encounter in real life,

and which usually require the use of oral language. Self-assessment and peer-assessment seemed to be starting to be used. We received six forms in which students assessed themselves and/or their classmates not just on their language development but also on their commitment to the subject and on their class participation. Finally, French teachers gave us three samples of papers (*trabajos*), i.e., a written work of about three pages or more, that required a search of information, be it bibliographical or through interviews, which was usually handed in at the end of a term and implied rather long-term planning. Examples of papers are research papers and final reports of project work.

Even though teachers in one of the programs used project work to assess students, they did not provide any written document with instructions or assessment criteria for the students. Neither did the teachers in the other program do so for portfolios, which have been promoted in that program. Only one teacher provided written guidelines to assess students' oral interviews.

In the following sections we analyze "hard" and "soft" assessment instruments in light of the characteristics of each.

4. Analysis of "hard" assessment instruments

According to Bachman and Palmer (1996), the most important consideration when designing a test is its usefulness, which is defined in terms of six qualities: reliability, construct validity, authenticity, interactiveness, impact, and practicality. None of these qualities should be disregarded at any of the other's expense, but a proper balance should be aimed at taking into account the purpose of the test, the characteristics of the domain of target language use (TLU) and of the test-takers, and the way the construct to be assessed has been defined. In this section these six qualities are used in order to analyze the instruments described above.

Reliability

Reliability relates to consistency of measurement. If someone takes the same test on two different occasions, she should have similar results. Also, if two

or more scorers give the same or similar scoring to a test, we can say that the test is reliable. Of course, this would only happen when the test-taker and test conditions are similar and when the scoring instructions are clear and specific. The lack of consistency of the scores when the test is taken under similar circumstances would cause the test to render unreliable results. "Of the many factors that can affect test performance, the characteristics of the test tasks are at least partly under control. Thus, in designing and developing language tests, we try to minimize variations in the test task characteristics that do not correspond to variations in TLU tasks." (Bachman and Palmer, 1996: 20–21)

Ever since testing emerged as a science (Piéron, 1969; Shohamy, 2001), reliability has been a major concern. With the aim of designing reliable tests, test designers have preferred objective items, i.e., items for which there is only one correct answer, such as multiple choice items like those used in well-known standardized tests. Reliability may be one of the reasons why teachers preferred to use objective items and not to use prompts. Objective items do not raise many arguments with students because of problems with scoring. Assessing productive skills such as speaking or writing is very subjective, and even if the teacher uses analytic scoring with clearly stated criteria, or invites another teacher to assess the same production—which takes more time—subjectivity will still be present. Besides, if the teacher uses global scoring she is liable to receive complaints from students who may not be satisfied with their grade. Therefore, in order to avoid trouble, teachers may avoid assessing writing. Another reason for using so many objective items in tests is that teachers do not need to spend much time in the design of tests . Either they take ready-made exercises from textbooks , or take a sentence or a text and leave out part of it for students to complete it. We shall discuss this further when we address practicality.

General instructions are important for transparency of assessment, because they are the means to inform test-takers how they are to proceed with the test, how the teacher will score it, and how the results will be used (Shohamy, 2001). Since no exams, quizzes, or drills had clear scoring procedures, reliability was at stake. If a student was unhappy with her test results, and a second scorer was called, the test results could be quite different. It might also happen that

the same scorer may score a test in a different way on two different occasions because there were no written scoring criteria.

Since no scoring procedure was specified**,** we may speculate that the teacher knew how she would score the test, and assumed students would know as well, or the teacher needed room for adjusting grades in case of surprises regarding students' responses. She might favor students if she gave less weight to a particular test part where very few students responded properly, but the opposite might also occur. If the scoring procedure has not been presented beforehand, students will have no support for possible complaints. This lack of explicitness shows the unequal power relations in the testing situation (Shohamy, 2001) and is related with opacity.

Construct validity

Construct validity is related to the significance and appropriateness of the interpretations made on the basis of test scores, and will depend on how we have defined the construct. The construct refers to the knowledge or skills that the test intends to assess. If we follow Bachman and Palmer (1996), knowledge of the language is defined in two areas: organizational knowledge and pragmatic knowledge. Organizational knowledge includes grammatical and textual knowledge. Grammatical knowledge refers to the ability to understand and produce utterances and sentences using accurate vocabulary, syntax, and phonology. Textual knowledge includes understanding and producing cohesive and coherent texts, i.e., understanding and producing explicit relationships between sentences and utterances and managing the organization of written texts and conversations.

Pragmatic knowledge, on the other hand, takes into account sociolinguistic knowledge and functional knowledge. Sociolinguistic knowledge allows us to understand and produce language that is appropriate according to the characteristics of the language use setting and covers the knowledge of dialects, varieties, registers, idiomatic expressions, cultural references, and figures of speech. Functional knowledge allows us to interpret relationships between utterances or sentences and texts, and the intentions of language users. The

functions may be ideational, manipulative, heuristic, or imaginative. The ideational function deals with the use of the language to express or exchange ideas, feelings, and knowledge. The manipulative function is intended to have things done, to control what other people do, and to establish, keep, and change interpersonal relationships. The heuristic function allows us to use the language to expand knowledge, for example, to teach or learn. Finally, the imaginative function permits us to use the language with aesthetic and humorous purposes such as writing poetry or understanding figurative language and jokes.

The majority of quizzes and workshops assessed only aspects of one of the components of organizational language knowledge: grammar and vocabulary, with items such as word completion, multiple choice, and transformation. This type of item makes it difficult to infer that the student is able to use the language in authentic situations. Exams were fewer and, besides grammar and vocabulary, had sections on listening and writing. However, since there was no indication about the criteria used to score the test, when there were writing tasks, for example, we do not know if it was grammar and spelling that were assessed or other features such as textual organization or appropriateness of register. If it was the latter, we think the task is more likely to have construct validity.

Regarding the pragmatic features of the language, most expected responses required an expression of an ideational function, very few required a heuristic function or an imaginative function. The language expected in responses was neutral, of standard variety with few cultural references and little figurative language. We did not observe much variation in the characteristics of the expected responses in the tests analyzed. If the tests were to evaluate students in level one or two, these characteristics would be appropriate, but since the expected response for upper levels has similar characteristics, we could think that the tests lack construct validity. Furthermore, if the purpose of these programs is to develop students' communicative competence, there seems to be a validity problem. It is difficult to infer that the student is able to communicate in the foreign language when quizzes and workshops carry more evaluative weight than exams.

Authenticity

For Bachman and Palmer authenticity is the "degree of correspondence of the characteristics of a given language test task to the features of a target language use task" (1996: 23). Authenticity is an important quality of tests for two reasons: it links students' performance in the test with TLU tasks and the domains in which the tester wants to use her interpretations. Besides, the way students perceive the relative authenticity of a test task may influence their performance on the test.

Most assessment instruments studied ranked low in authenticity since in most tasks students were to select the right answer or to complete sentences with a word or a phrase—tasks that are rare in situations other than exams. Prompts, whereby students are asked to elicit an extended response, like the ones we produce in real-life language use, were scarce. However, we found a "listening quiz" where students had to watch a movie and comment on it (even though these comments were written). Furthermore, the teachers' focus was generally not on performance on tasks similar to TLU tasks but on the four skills and their sub-skills. Exams had different sections, such as vocabulary, grammar, listening, reading, and/or writing; which were usually presented in this order. They were not organized based on what students are able to do in particular situations. Thus, from tests it is difficult to infer how the student will perform in real situations, which may be one reason why some students in upper levels are not able to do what they are expected to do with the language. They may have received good grades on each test on grammar, on vocabulary, on reading, etc., but they do not know how to deal with particular situations.

It is believed that the more authentic a teaching situation, the better and more authentic the learning. Wilson (1995, in Shohamy, 2001) states that "language is learned in use; language use is context related. Language evaluation therefore must occur in authentic contexts" (171). Both programs advocate for the foreign language to be learned by using it in context; however, many of the instruments provided by participants lacked authenticity; few had a real life approach. There seems to be no coherence between how students are taught and how they are assessed. Teachers may be using more in-class tasks or

projects, which are by nature more authentic than tests. However, they do not keep written records of the instructions they give, nor the criteria they use to assess students.

Interactiveness

Bachman and Palmer (1996) characterize interactiveness of a language test task as the way in which the test-taker's language ability, topical knowledge, and affective schemata are engaged in the test task. Language ability involves using the knowledge of the language—divided in the two areas mentioned above—and strategic competence. Strategic competence involves using executive processes to regulate the cognitive functions engaged in solving the test tasks. When the student uses this set of meta-cognitive strategies she decides what she will do, analyzes what she needs to solve the test task and the possibilities she has to complete the task satisfactorily, and decides how she will use the knowledge she possesses.

Topical knowledge affects students' response especially in reading and listening tasks. If the students are familiar with the topic, they have a better chance to succeed in the task. Similarly, affective schemata influence students' responses. There may be certain topics that attract students and other topics to which they are quite sensitive. The length and the layout may also encourage or discourage test-takers. Since interactiveness depends on the characteristics of the students, any analysis we do of test tasks will only be tentative, since we do not know the students.

Most tests that included objective items assessed grammar and/or vocabulary and thus ranked low in interactiveness because tasks did not require much use of students' meta-cognitive strategies. Also, the language knowledge shown was very restricted, making it difficult to infer how students will perform in TLU tasks. Regarding their knowledge of the topic and their affective schemata, tasks were quite neutral, because most dealt with personal topics that are not sensitive. In reading tasks included in tests the topics were also personal, with some exceptions such as cultural or technical topics, which we think did not affect the students' responses because the questions asked were very direct,

i.e., they required mainly literal comprehension of specific parts of the text. However, we found a text which described Miami and had lots of figures, which we think might negatively affect those students who feel estranged from that cultural content and/or might dislike numbers.

Regarding the layout, we feel that most instruments had an attractive appearance, which may encourage students to respond to them. However, we found tests in which completion items were written, one after the other, as if they were text. We feel that this might annoy some students who will try to find coherence among the sentences and fail to do so. Single-spaced reading texts in small font may have a similar affective reaction on the part of test-takers.

Impact

Impact is the last quality related to use and interpretation of tests. The impact of a test refers to the way its results affect individuals, institutions, or society. Topical knowledge, familiarity with testing techniques and instructions, input, the feedback given, and the decisions made on the basis of the test results are some examples of how test scores might affect individuals positively or negatively.

It is difficult to ascertain the impact of the assessment instruments at the micro- and macro-level, i.e., how they affect individuals and how they affect the educational system or society at large. We know that tests are used to determine if students pass or fail the course; however, tests are not the only kind of information teachers use to make this decision. We will discuss this further when we address "soft" assessment.

Test scores give students feedback on their performance and promote introspection on the learning strategies they are using and on their commitment, for example. Besides, students whose teachers provide more descriptive feedback on their test results—either in whole class or in individual advisory sessions—may be better able to introspect and improve. Furthermore, students might use the scores to make the decision to quit the course for fear of failing; otherwise they would be excluded from the program or have to pay in order to repeat the course.

Exams, quizzes**,** and the rest of the instruments were, to some extent, tailored to what had been taught, rather than the other way around. Since the individual teacher designed the assessment, we would expect a positive washback effect on their teaching and on students' learning, especially when feedback was given.

Practicality

Practicality, the quality dealing with implementation, is defined by Bachman and Palmer as "the relationship between resources that will be required in the design, development, and use of the test and the resources that will be available for these activities" (1996: 36). This means that human, material, and time resources are essential to estimate the practicality of a test. Thus "a practical test is one whose design, development, and use do not require more resources than are available" (Bachman and Palmer, 1996: 36). Practicality seemed to be the criterion that most influenced how teachers assessed their students. Since teachers have to move from one institution to another and are in charge of so many students, time to design, write, administer, score, and analyze the results of the assessment instruments has to be reduced. This may be the reason why objective items in which students select their response or in which they complete with a word or phrase are preferred. These types of tests may be constructed by photocopying exercises or taking items from textbooks or other tests and are easy to administer. They do not take much time to correct either, a crucial factor to consider when we think that all teachers in these programs are hired by the hour and work in different institutions. As one teacher stated:

> Quote 3: I try like not to make a variety of exercises. It should have three or four points. It should be one exercise, just one activity and that's it. A reading comprehension exercise. (…) I try not to mix (…) because of time concerns (…) in order not to make it too complicated to mark, because it is easier to mark an exercise where there are ten multiple choice items than an exercise where they have to complete with vocabulary they have learnt, where there is a reading and they have to include certain things, or reading comprehension. (Interview with Teacher 9)

Time may also affect the input of language tests, which was mainly written. Only few teachers assessed listening using audiotaped texts or videos, probably

because finding the appropriate materials is also time consuming. The lack of use of other channels in French may be explained by the scarce availability of resources compared to English, which means that teachers go out of their way to find audiotaped texts or videos.

This idea of practicality may also be influencing the fact that teachers did not assess students in writing tasks because that would mean too much correction and work. One of the teachers referred to this in an interview:

> Quote 4: Sometimes one does not have the time to sit down and say "Well, I am going to correct this at once." Time is not enough. You get home at nine or ten o'clock at night and don't want to see any more papers at least until eleven. So that is the difficult part. That is why I am against writing tests. They are the ones I do the least. (Interview with Teacher 1)

In this section we have used Bachman and Palmer's qualities of tests to do the analysis of "hard" assessment instruments; in the following section we analyze "soft" assessment instruments using these qualities and those advocated by supporters of "soft" assessment.

5. Analysis of "soft" assessment instruments

Besides the qualities of useful language tests mentioned by Bachman and Palmer (1996) there are other characteristics that should be considered in "soft" assessment: democracy in the design, administration and interpretation of results, fairness and transparency. Furthermore, proponents of "soft" assessment (Bustos, 1997; Jorba and Casellas, 1997) value its formative character and its authenticity. On the whole, the purpose of portfolios and self- and peer-assessment is to make students critically aware of their own learning process; since it assesses it in an ongoing manner, the students receive timely feedback on their performance. On the other hand, the purpose of papers, in-class tasks and role-plays is to provide students with more authentic alternatives to show their progress in language learning. Because they are carried out on a regular basis to assess everyday activities, they focus on both process and product. Moreover, these alternatives in assessment require students to make use of

their creativity to produce or do something; therefore, they require problem solving and thinking skills similar to those required in TLU tasks. Among the advantages of this type of assessment, we find that it is more personalized and thus, individual differences are more likely to be taken into account, making this type of assessment more flexible and fairer.

Despite all these advantages, there were few instances of "soft" assessment instruments. Devising written instructions for role-plays or portfolios is time-consuming and (if we want to be democratic) reaching consensus with students on grading criteria is even more time-consuming. Correcting portfolios periodically, revising papers, and giving feedback is also tedious and difficult work for teachers who are hired by the hour and are always on the run. This relative difficulty concerning organization and production—which are related to practicality—and subjectivity in grading are considered as some of the disadvantages of using "soft" assessment by Brown and Hudson (1998).

Portfolios are considered "a collection of students' work that demonstrates to students and others their efforts, progress and achievements in given areas" (Genesee and Upshur, 1996:99). However, in our analysis of instruments, we found that there seems to exist some confusion because the sample of student portfolios handed in by one of the participating teachers included both a mixture of quizzes and drills plus some samples of "soft" assessment without clear organization. There were no traces of feedback given to the student, so it seemed to have a more summative than formative purpose.

Even though portfolios and self- and peer-assessment are not primarily meant to be used for grading, in many cases they are also used for making decisions about students' passing or failing. This may be the reason why Norris et al. (1998) claim that "the issues of reliability and validity must be dealt with for alternative assessments just as they are for any other type of assessment—in an open, honest, clear, demonstrable, and convincing way" (Brown and Hudson, 1998: 5). Most instruments used by teachers in in-class tasks had neither explicit instructions regarding time for preparing and performing the task, nor the specified criteria for good performance. The samples of students' papers and the portfolio did not provide any information on criteria for grading them

either. This affects their reliability and their construct validity. It also affects their transparency.

In general, we consider that in-class tasks and papers were highly authentic and interactive. For example, in an in-class task students had to try to organize a group of people for a photo, a task similar to what we do in real life. This task requires students to use their strategic competence and makes it easier for the teachers to infer if students have the knowledge of the language required for performing in a similar situation out of class.

Regarding the impact of assessment, besides taking tests into account, teachers also take into consideration students' use of the language in class activities or in oral interviews, and affective variables such as motivation, commitment, and effort.

> Quote 5: Sometimes I start counting, he has more "+", he participated, he was often in class, he researched. So, for example, if I see he is kind of weak, but he has good participation, I pass him. (Interview with teacher 12)

In both programs, there were some teachers who promoted self- and peer-assessment, which makes assessment more democratic by allowing students to get involved in the process, and fairer by having other sources of information other than tests. The forms for self- and peer-assessment required students to rate their own language knowledge and their commitment, so they also helped students to understand the meaning of learning and promoted their autonomy, which constitutes part of the philosophy in both programs, thus promoting a more holistic education.

> Quote 6: I also tell students they have to self-assess and I emphasize that role a lot so that they become more participative. (I like) students to assess themselves and to express it at the end of the course, or during the course. And I feel satisfied when they feel more as persons than as elements in a class. (Interview with Teacher 10)

Even though alternatives in assessment were being implemented by participant teachers and tasks ranked high in authenticity and interactiveness, we found that

qualities such as construct validity, reliability, and transparency were affected by the lack of clearly explicit criteria and procedures. The few instances of "soft" assessment instruments that were provided by teachers and the fact that they had no traces of how they were used, make it difficult to determine if qualities of "soft" assessment such as its formative character, fairness and transparency were fully exploited. However, we think that encouraging self- and peer-assessment among students gives the process a more democratic character.

Conclusion

Assessment in the teaching of a foreign language becomes paramount. Many complex aspects are involved: students' characteristics—their knowledge of the language, knowledge of the topic, strategic competence, and affective schemata—the domain of TLU, and the working and social characteristics of the teachers.

In this study, the instruments teachers use to assess their students are mainly of a "hard" type, like quizzes, exams, and drills. Few teachers provided instruments for "soft" assessment. Apparently, the qualities of assessment that teachers cherish the most are practicality and reliability, and the ones least taken into consideration are authenticity and interactiveness. The reasons for this are probably the lack of available time for teachers, which makes them look for objective items which are easy to correct and shun the design and assessment of more naturalistic tasks. Construct validity is a problematic issue since most instruments assess similar characteristics of the language, mainly basic vocabulary and grammar, and do not specify the kind of tasks the students are to perform in order to show their competence. All this makes test scores difficult to interpret.

"Soft" assessment is starting to be used, however, but we do not see clarity regarding the definition of the construct and the criteria used to assess students. At the moment of deciding who passes and who fails a course what does the teacher take more into account: knowledge or commitment; her opinion of

the student, the student's own opinion, or students' view of their peer? The following stage in this research process, namely, the analysis of teachers' and students' interviews, may throw more light on this complex issue.

References

Alte Members, 1998, *Mutilingual Glossary of Language Testing Terms***,** Cambridge, Cambridge University Press.

Altrichter, H., P. Posch, and B. Somekh, 1993, *Teachers investigate their work. An introduction to the methods of action research*, London, Routledge.

Bachman, E.L., and A.S., Palmer, 1996, *Language Testing in Practice: Designing and Developing Useful Language Tests*, Oxford, Oxford University Press.

Brown, J. D., and T. Hudson, 1998, "The Alternatives in Language Assessment", *TESOL Quarterly*, 32(4), Waldorf.

Burns, A., 1999, *Collaborative Action Research for English Language Teachers*. Cambridge, Cambridge University Press.

Bustos, F., 1997, *Mutaciones en Evaluación y Lineamientos del Ministerio: Análisis y Crítica.* (Mimeo excerpt published by Vicerrectoría de Docencia, Universidad de Antioquia).

Carroll, B. J., 1993 (aoû.-sep.), "Typologie des tests de langue", *Le Français dans le Monde,* numéro spécial: Évaluation et certifications en langue étrangère.

Genesee, F., and J.A. Upshur, 1996, *Classroom-based Evaluation in Second Language Education,* Cambridge, Cambridge University Press.

Instituto Cervantes, 2002, *Marco de referencia europeo para el aprendizaje, la enseñanza y la evaluación de lenguas*, (Translation by Alejandro Valero Fernández), Web site: http://cvc.cervantes.es/obref/marco.

Jorba, J., and E. Casellas, eds., 1997, *La regulación y la autorregulación de los aprendizajes,* Madrid, Síntesis.

Kemmis, S. and R. McTaggart, 1988, *The Action Research Planner*, Australia, Deakin University Press.

Piéron, H., 1969, *Examens et docimologie*, Paris, Presses universitaires de France.

Salinas, M.L., s.f., *La Evaluación de los Aprendizajes en la Universidad*, Facultad de Educación, Universidad de Antioquia

Shohamy, E., 2001, *The Power of Tests. A Critical Perspective on the Uses of Language Tests*, Essex, Pearson Education Limited.

The authors

** Maestría en Educación Especialidad: Lingüística Aplicada. Miembro del GIAE (Grupo de Investigación Acción y Evaluación en Lenguas Extranjeras).
Correo electrónico: cfrodden@hotmail.com

*** Estudiante de la Escuela de Idiomas Universidad de Antioquia. Pregrado en curso de Lenguas Extranjeras. Miembro del GIAE (Grupo de Investigación Acción y Evaluación en Lenguas Extranjeras).
Correo electrónico: srestrepo@idiomas.udea.edu.co

**** Especialización en Lenguas Extranjeras –Enseñanza Docente Auxiliar-Tiempo Completo. Fundación Universitaria Luis Amigó – Vicerrectoría Académica – Departamento de Idiomas. Miembro del GIAE (Grupo de Investigación Acción y Evaluación en Lenguas Extranjeras).
Correo electrónico: lilith39@epm.net.co

Annex 1

Quote 1: *El profesor colombiano, en el caso por ejemplo del profesor de cátedra, tiene tanto que correr y tiene que ir de una institución a otra, y tiene que esforzarse tanto por hacer un salario decente para pagar sus cuentas que no tiene el más mínimo ánimo de llegar a la noche a escribir.*

Quote 2: *Trato de que cada examen tenga contenido de una estrategia de evaluación diferente. Entonces, trato de que un punto sea apareamiento (sic), que otro sea completación, que otro punto sea de escritura, que otro punto sea de comprensión auditiva, o un dictado o que sea también, por ejemplo, de falso o verdadero, es decir, que sea lo más dinámico posible.*

Quote 3: *Yo trato de pronto de hacer un ejercicio no muy variado. Que sea de tres o cuatro puntos. Sea que un ejercicio, una sola actividad y punto. Una comprensión de lectura. (...) Trato de no mezclar (...) por cuestión de tiempo. (...) para de pronto no complicarme al momento de calificar porque es más fácil calificar un ejercicio donde hay diez puntos de selección, a un ejercicio donde hay selección múltiple, donde tienen que completar vocabulario ya conocido, donde hay una lectura y tienen que incluir ciertas cosas, o comprensión de lectura.*

Quote 4: *A veces no tiene uno como tiempo de sentarse y decir "bueno, me voy a sentar a corregir esto ya". No le da, llega uno a la casa a las nueve o diez de la noche y no quiere ver más papeles por lo menos hasta las once; entonces, esa es la parte difícil. Por eso, también soy enemigo de hacer exámenes de escritura, son los que menos hago.*

Quote 5: *Hay veces que yo, bueno, empiezo a contar, tiene más "+", me participó, vino mucho en clase (sic), me consultó, entonces ya por ejemplo, si veo que como que está flojo, pero tiene buena participación yo, yo lo paso.*

Quote 6: *También les digo a ellos, pues, deben autoevaluarse y enfatizarles mucho siempre (...) ese papel para que sean más participativos. Que los estudiantes se evalúen y lo manifiesten al final también de los cursos o durante los cursos. Y me da satisfacción donde ellos se sienten más como personas que como elementos dentro de la clase.*

Annex 2

Chart to analyze assessment instruments

Based on concepts taken from: Alte Members (1998); Instituto Cervantes (2002); and Bachman and Palmer (1996).

TEACHER

INSTITUTION, PROGRAM, LANGUAGE, AND LEVEL

TYPE OF ASSESSMENT

Global	Non-global/Discrete point

SCORING

Global	Analytic/Discrete point
Subjective	Objective

NAME OF THE INSTRUMENT

Exam / Test	Quiz	
Drill(s)	In-class task	Portfolio
Project	Paper	Other:

RUBRIC

Instructions			
Language: Mother tongue Foreign			
Channel: Aural Visual Other Specify:			
Specific instructions for: General procedures: Yes __ No ___ Part(s): Yes __ No ___			
Structure			
Number of parts:		Number of items per part:	
Types of Items			
Multiple choice	True or false		Matching
Sentence completion	Word completion		Cloze-type of exercises
Unscrambling words	Reordering of texts		Transformation
Closed question	Open question		
Oral or written production with specified topic		Oral or written production without specified topic	
Dictation		Other Specify:	

CHARACTERISTICS OF THE INPUT

<table>
<tr><td colspan="2">Format</td></tr>
<tr><td colspan="2">⇒ Channel of presentation</td></tr>
<tr><td colspan="2">Printed: Text Graph/Picture Screen</td></tr>
<tr><td colspan="2">Aural: Recorded: Audio Audiovisual Live:</td></tr>
<tr><td colspan="2">⇒ Length: Lines: Minutes:</td></tr>
<tr><td colspan="2">⇒ Type: Item: Prompt:</td></tr>
<tr><td colspan="2">⇒ Language: Mother tongue: Foreign: Both:</td></tr>
<tr><td colspan="2"></td></tr>
<tr><td colspan="2">Characteristics of the language</td></tr>
<tr><td colspan="2">Organization</td></tr>
<tr><td>Grammatical</td><td>Textual</td></tr>
<tr><td>♣Vocabulary:
• Restricted
• Broad</td><td rowspan="4">♣Expository
♣Descriptive
♣Narrative
♣Argumentative</td></tr>
<tr><td>♣Morphosyntax
Length of sentences
• Short (max. 10 words)
• Intermediate
• Long (more than 2 lines)
Structure
• Only S + V (+ …)
• With Adverbial + S + V (+ …)
Degree of control
• Controlled (textbook-type)
• Free (authentic)</td></tr>
<tr><td>♣Written presentation or graphology
• Printed
• Hand-written
• Quality: clear blurry fragmented</td></tr>
<tr><td>♣Phonology
In the case of cassettes, CDs, videos, etc.</td></tr>
<tr><td colspan="2">Pragmatic</td></tr>
<tr><td>Functions</td><td>Sociolinguistic aspects</td></tr>
<tr><td>♣Ideational
♣Manipulative
♣Heuristic
♣Imaginative</td><td>♣Language: Standard Variety
♣Register: Casual Neutral Formal
♣Cultural references
♣Figures of speech</td></tr>
<tr><td>Topic</td><td></td></tr>
<tr><td>Academic</td><td>Personal</td></tr>
<tr><td>Technical</td><td>Cultural</td></tr>
</table>

STUDENTS´ EXPECTED RESPONSE

<table>
<tr><td colspan="2">Characteristics of the language</td></tr>
<tr><td colspan="2">Organization</td></tr>
<tr><td>Grammatical</td><td>Textual</td></tr>
<tr><td>♣Vocabulary:
• Restricted
• Broad</td><td rowspan="4">♣Expository

♣Descriptive

♣Narrative

♣Argumentative</td></tr>
<tr><td>♣Morpho-syntax
Length of sentences
• Short (max. 10 words)
• Intermediate
• Long (more than 2 lines)
Structure
• Only S + V (+ …)
• With Adverbial + S + V (+ …)
Degree of control
• Controlled (textbook-type)
• Free (authentic)</td></tr>
<tr><td>♣Written presentation or graphology
• Printed
• Hand-written
• Quality: clear blurry fragmented</td></tr>
<tr><td>♣Phonology
In the case of cassettes, CDs, videos, etc.
.</td></tr>
<tr><td colspan="2">Pragmatic</td></tr>
<tr><td>Functions</td><td>Sociolinguistic aspects</td></tr>
<tr><td>♣Ideational
♣Manipulative
♣Heuristic
♣Imaginative</td><td>♣Language: Standard Variety
♣Register: Casual Neutral Formal
♣Cultural references
♣Figures of speech</td></tr>
</table>

Topic	
Academic	Personal
Technical	Cultural

Annex 3

(Name of the university)
(Name of the school)
(Name of the program)

Written tests for true colors 1 / test 1: units 1–5

(Don't write on this test, write only on the answer sheets)

1. COMPLETION

Write the correct word in the blank.

1.1. My brother _____ a new job. (have / has)
1.2. _____ your parents have a car? (do / does)
1.3. Sarah _____ like rock music. (don't / doesn't)
1.4. John and Lisa _____ part-time. (work / works)
1.5. Brett doesn't _____ to school. (go / goes)

2. ELABORATION

Complete the sentences with activities from the box. Use the Present Continuous and contractions.

~~Talk to a friend~~	cook dinner	play soccer	work	brush teeth

Example: Jason is on the phone. He ***'s talking to a friend.***

2.1. Bob's in the kitchen. He
_____

2.2. Sally's at her office. She
_____

2.3. Jane and Tina are in the living room. They
_____

2.4. Sam and Eric are outside. They
_____

2.5. Pat's in the bathroom. She
_____

3. QUESTIONS AND ANSWERS

3.1. Answer these questions:

3.1.1. What do you cook for dinner?
3.1.2. When does your best friend go shopping?
3.1.3. How many letters do you write a month?
3.1.4. What kind of parties do you go to?
3.1.5. Which ice cream do you buy at the supermarket?

3.2. Create logical questions for these answers:

3.2.1.
_____?
Because I'm having lunch right now!
3.2.2.
_____?
His father works in an office.
3.2.3.
_____?
Ellen eats breakfast at seven o'clock.
3.2.4.
_____?
She exercises every morning.
3.2.5. _____?
It never snows in Medellín!

4. THE RIGHT DESCRIPTION

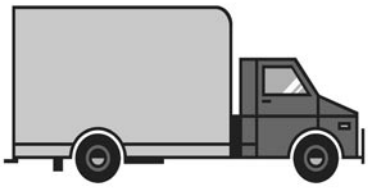
Write the most appropriate description for the following pictures, using **3** ***descriptive adjectives.***

4.1 Picture 1

4.2 Picture 2

4.3 Picture 3

4.4 Picture 4



4.5 Picture 5

5. GRAMMAR KNOWLEDGE Write the correct word in the blank.

5.1. Is Bill in _____ room?

a. his
b. its
c. he

5.2. _____ name is Andrea. I'm from Chicago.

a. I
b. Your
c. My

5.3. Could you spell _____ name, please?

a. your
b. our
c. yours

5.4. Let's go to a restaurant tonight, honey. It's _____ wedding anniversary.

a. your
b. our
c. their

5.5. That's a present for my son. It's _____ birthday tomorrow.

a. its
b. his
c. her

6. MATCH	Match the questions and the answers.
6.1. Are you studying computers?	a. No, she isn't.
6.2. Does Ann work full-time?	b. No, they aren't.
6.3. Are they students?	c. Yes, I do.
6.4. Do they work full-time?	d. No, they don't.
6.5. Does he like his job?	e. No, she doesn't.
6.6. Do you like reggae music?	f. Yes, we do.
6.7. Do you two eat out on weekends?	g. Yes, he does.
6.8. Does she like you?	h. Yes, she does.
6.9. Do I look good in green?	i. No, I'm not. I'm studying music.
6.10.Is your mother cooking for you now?	j. Yes, you do.

7. DICTATION Listen to what the teacher is going to dictate for you:

7.1. _____

7.2. _____

7.3. _____

7.4. _____

7.5. _____

Reprinted with the author's permission.