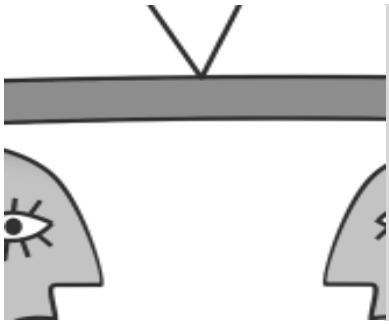
REVISTA Universidad EAFIT
Vol. 39. No. 132. 2003. pp. 65-75

Estimating the validity
and reliability of an oral assessment instrument



Ana Patricia Muñoz Restrepo
M.A. in TESOL, *Eastern Michigan University*. Coordinadora de Investigación y Docencia del Centro de Idiomas de la Universidad EAFIT.
apmunoz@eafit.edu.co

Martha Eugenia Álvarez Villa
Ingeniera Industrial, Especialista en Sistemas de Información. Profesora del Departamento de Ingeniería de Sistemas de la Universidad EAFIT.
ealvarez@eafit.edu.co

Recepción: 27 de enero de 2003 I Aceptación: 12 de marzo de 2003

Abstract
This article reports partial findings on a study conducted to estimate the validity and reliability of an oral assessment instrument.The study was conducted during the year 2002 at the Language Center in Eafit University. Validity was determined both through logical and empirical procedures. The results indicated the need to define clearly and precisely the construct to be measured, specify assessment criteria and scoring, and have evaluators who thoroughly understand the aspects included in the assessment instrument. Reliability was estimated using Spearman rank correlations as well as One Way Anova. Results suggest that well trained evaluators and well-designed instruments provide high consistencies in measurement.

Key Words
Confiabilidad / validez / competencia comunicativa

Resumen
Este artículo presenta los resultados parciales de un estudio realizado en el año 2002 para determinar la validez y confiabilidad de un instrumento para medir el desempeño oral de estudiantes de Inglés en el Centro de Idiomas de la Universidad Eafit. La validez se estimó mediante procesos empíricos y lógicos. Los resultados indican la necesidad de definir claramente la habilidad que se quiere medir, especificar los criterios y puntajes de evaluación y contar con evaluadores que comprendan adecuadamente los aspectos a medir. La confiabilidad se midió a través de correlaciones y análisis de varianza. Los resultados sugieren que es posible obtener consistencias en los puntajes de varios evaluadores cuando la herramienta de evaluación es bien diseñada y cuando se ha entrenado a los evaluadores.

Palabras claves
Reliability / validity / oral language ability

Introduction

he approach to language teaching used at the EAFIT Language Center is the communicative approach, which concentrates on developing the learner's ability to communicate effectively through meaningful and authentic situations for the learner. Grammar study is viewed as just one of the vehicles that can be used to promote communicative competence (Flaitz, 2000:4). Under this approach, instructors make use of authentic, or real-life, situations and activities that require communication and that are relevant to the lives of the learners –- role-plays, games, interviews, problem solving activities, and the like. The Language Center's communicative approach to teaching and learning is also present in its assessment practices. It proposes the assessment of spoken language through:

- A variety of tasks aiming at different learning differences
- Authentic and meaningful tasks
- Different grouping techniques to elicit interaction among the students and with the teacher
- Encouragement of self and peer assessment
- Assessment tasks derived from curriculum objectives and consistent with instructional practices
- Ongoing assessments so that students can demonstrate the extent of their knowledge and abilities
- Assessment of different aspects of oral language where grammar is only one of many different aspects considered in the assessment of communicative competence

The congruence between methodology and assessment practices, as well as the use of fairly valid and reliable assessment tools, is essential in establishing desired teaching practices. In other words, the agreement between methodology and assessment may bring positive washback in the classrooms.

With this philosophy in mind, teachers are encouraged to view assessment as an integral part of language learning and teaching. Sometimes assessment is used for the teacher to assess his/her own effectiveness in teaching the goals, and it is also needed to see where students are in relation to the goals. At other times, it is used to provide additional assistance to students who are struggling with certain concepts.

Nonetheless, assessment, more precisely, *oral assessment* is a challenging endeavor given (i) the different teaching practices and beliefs each teacher has, (ii) the lack of specific assessment criteria (tools and standards), and (iii) the lack of systematic and ongoing procedures. This leads the teacher to subjective and impressionistic assessments, which are clearly a central part of language assessment, but which do not reveal teachers' basic understanding of the principles of assessment and of the ability being measured. These difficulties, however, may be overcome and a consensus on similar assessment and feedback practices achieved through teacher training and the appropriate use of assessment tools. Put differently, it is possible, through training, discussions, and valid and reliable assessment tools to decrease the variability involved in the assessment of oral language.

1. Background

The investigation of assessment practices at the Language center came from a 1999 research study conducted at the Language Center which focused on three areas: 1. The effectiveness of the required materials in reaching the Language Center's oral goals, 2. the students' and teachers' beliefs about the role of oral language in the classroom, and 3. oral assessment. What the research project on the area of assessment showed was that more investigation needed to be done into what was actually happening in the classroom and to gather more information on specific materials and methods being used (Cohen and Fass, 2001).

To this end, a 2001 study looked into teachers' beliefs and practices in assessing spoken language (Muñoz, et al., 2003a). Namely, teachers were asked about methods, materials, aspects of oral language, frequency, and reasons for doing assessment. The results indicated that most teachers focused assessment on summative purposes. Very few teachers considered assessment a process through which teaching methodology and learning can be improved. A lack of systematic and ongoing procedures in assessment was also revealed, being this an indication of teachers' unawareness of the goals of assessment and lack of planning when assessing students.

The implications of this study were: 1) the need for educational programs in the area of assessment, and 2) the development of an oral assessment instrument that would allow teachers to have a consensus on similar assessment and feedback practices.

In 2002, a tool for assessing oral language was developed aiming at its implementation in the Adult English program. The instrument, called Oral Assessment System (OAS), consisted of:

- A rubric for oral assessment
- An oral assessment grade sheet to keep record of assessment activities and grades
- A report card to inform students of mid - term and final evaluations

The step that followed was to determine if the system was useful. The usefulness of an assessment instrument can be examined by looking into two test qualities: validity and reliability (Bachman and Palmer, 1996). This article presents the procedures followed in estimating the validity and reliability of the Oral Assessment System.

2. Theoretical Framework

In the design of any assessment instrument test developers must be concerned with 1) identifying potential sources of error in the instrument and 2) providing evidence to justify test score interpretations. These two aspects will be addressed and discussed as reliability and validity respectively.

Reliability consists of estimating the amount of variation in language test scores that is due to measurement error (Bachman, 1990). This is crucial in language assessment because students' performance on a test may be affected by factors other than the ability that we are measuring (illness, fatigue, poor test conditions, poor test design, score inconsistencies). In oral language assessment, we are specifically concerned with sources of error due to inconsistencies among raters (inter rater reliability) because of the subjectivity involved in measuring an abstract entity.

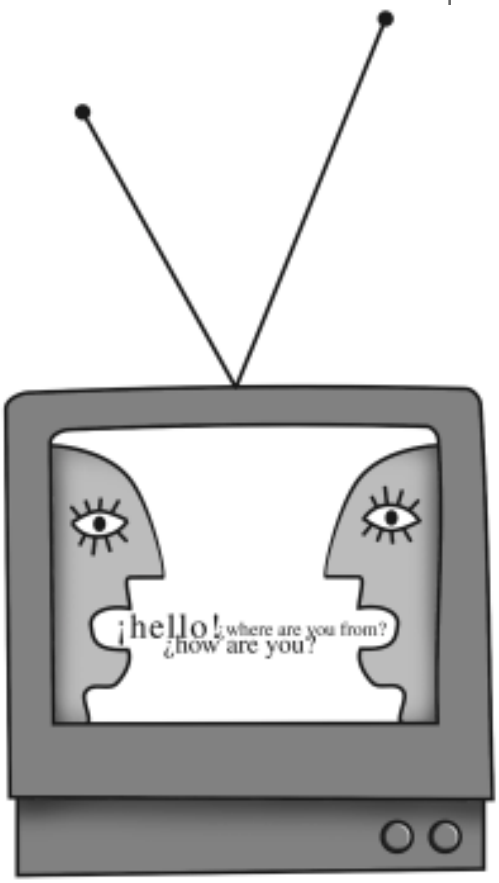
Indeed, the reliability of oral assessment has been a critical issue in testing research. Seward (1973: 76), for instance, claims that oral tests "are impressions of the tester about the student's speaking ability rather than accurate objective measures of speaking proficiency." Ingenkamp and Wolf (1982:341) report a "remarkable between-rater variance" in a study of oral examinations for the German *Abitur* (matriculation exams). This is obviously natural when evaluating aspects that are very hard to define, such as communicative effectiveness, fluency, etc. For this reason, the increasing use of subjective assessments leads to a corresponding need to establish the reliability and validity of such assessments.

A way to improve inconsistencies due to subjective ratings is to use more than one evaluator. Variance among different raters can also be improved by reaching a consensus through training and discussions and by establishing clear oral performance criteria and scoring. Criteria and scoring act as guidelines for judgment that should clearly describe the various levels of performance in a way that can be tested both logically and consistently (validity and reliability). In addition to the specificity of criteria and scoring, the manner in which raters are trained also plays an integral role in determining consistency within the scoring process (Herman, Aschbacher, & Winters, 1992).

According to Bachman, (1990:160) when we increase reliability "we are also satisfying a necessary condition for validity: in order for a test to be valid, it must be reliable." This provides a framework from which reliability and validity can be interpreted, not

as separate test qualities, but as interrelated. As Bachman (1990) points out, they both are concerned with identifying, estimating, and controlling the effects of factors that affect test scores.

Validity has been defined as the extent to which a test (or any assessment instrument) truly represents the ability we want to measure. The question here is, however, how to describe the ability that we want to measure when it comes to second or foreign language learning. As Jakobovits (1970:95) argues "what is to know a language is not yet well understood and consequently the language proficiency tests now available […] attempt to measure something that has not been well defined." Furthermore, how can we justify the interpretations we make of a score as an indication of student's language ability?



To justify interpretations of a test score, it is necessary to provide evidence that the score represents the language ability that we want to measure. It is then necessary, first of all, to provide a clear definition of the ability or construct to be measured. In our specific situation, we are concerned with describing the construct *communicative competence* or *language ability*. This is extremely important because communicative competence determines not only what instruction a student receives, but also what instruments teachers will use to measure language progress over time. Second, when assessment is founded on a clear theory of language use, it is possible to choose relevant assessment tasks.

The definition of the construct provides a means for investigating validity. According to Messick (1989), the unified validity of a test is revealed through an overall evaluative judgment of the instrument. This judgement requires a comprehensive evaluation based on theoretical rationales and empirical evidence that support the adequacy and appropriateness of interpretations and actions based on test scores.

From an empirical point of view, validity of a language test can be examined by looking at the correlation between students' scores on the test being examined to other scores from a similar and already

validated test (concurrent validity). Such evidence provides a way for understanding the utility or meaning of test scores. Correlational studies have commonly been used in language testing to examine patterns of correlations among test scores, either directly, or, for correlations among large numbers of test scores, through factor analysis (Bachman, 1990). High correlations are taken to indicate that two tests measure the same language ability while low correlations suggest this is not the case.

3. The Validity of the Oral Assessment System

The development of the oral assessment instrument, called Oral Assessment System (OAS) began by deciding which type of instrument should be used to fit the needs of teachers, students, and the institution. The researchers set out to develop a rubric (set of scoring guidelines for evaluating student work) for oral assessment. A rubric provides for increased consistency in the rating of performances and understanding by giving students an established set of expectations about what will be assessed as well as the standards that need to be met. A rubric is an authentic assessment tool, which is particularly useful in assessing criteria that are complex and subjective. A rubric provides several advantages**:** 1) allow assessment to be more objective and consistent; 2) focus the teacher to clarify his/her criteria in specific terms; 3) clearly show the student how their work will be evaluated and what is expected; 4) promote student awareness of the criteria to use in assessing self and peer performance; 5) provide useful feedback regarding the effectiveness of the instruction. A typical rubric contains:

- A scale of possible points to be assigned in scoring work.
- Descriptors for each level of performance
- Aspects of language to be assessed

The design of the rubric began by considering the oral language aspects mentioned by teachers in the research study (2001) mentioned above. Specifically, these aspects were: fluency, pronunciation, grammar, understanding, comprehensibility, and vocabulary. With these aspects in mind, the researchers developed a rubric for oral assessment.

To determine if the OAS was a valid instrument logical and empirical analyses were conducted. A logical analysis implies defining theoretically the ability or construct to be measured. In our situation this construct refers to "communicative language competence" or language ability.

However, defining language ability is not an easy task especially when it comes to language learning. For instance, what is second language aptitude? What is intelligence? What is communicative competence? These questions have been subject to debate in applied linguistics and will continue to be. Nonetheless, we can offer a fairly comprehensive description of language competence. In so doing we followed theoretical models described by different authors who have worked on providing a better understanding of communicative competence over the past 20 years (Canale and Sawain, 1980; Savignon, 1983; Bachman, 1990; Bachman and Palmer, 1996). In our model, being communicatively competent requires more than learning the grammatical and lexical components of language. Communicative competence is demonstrated through the ability to communicate and negotiate meaning by interacting meaningfully and accurately with other speakers. Therefore our model is based on linguistic, socio-linguistic, discourse and strategic competences. A broad description of this model may be found in "Guidelines for Oral Assessment," a working paper used in in-service training programs for the Language Center teachers. The definition of language competence allowed us to establish the kinds of behaviors we wanted to observe in our students in terms of oral language. These behaviors or performance criteria needed to be reflected in the assessment instrument.

Based on the definition of communicative competence, the 2001 rubric was redesigned in order to have a better representation of the construct. The 2001 rubric only covered linguistic aspects of communicative competence, which meant that the construct was being 'under-represented' and therefore threatening validity (Messick, 1996). The logical analysis of validity led to multiple modifications and improvements in the rubric. Basically, sociolinguistic,

discourse, and strategic competences were added in order to have a more comprehensive representation of oral language ability.

In the revised rubric changes were also made in the scoring system. The 2001 rubric included a letter grade scale A, B, C, D, interpreted as follows: A= Excellent; B= Good; C= Fair; D= Fail. These letter grades were substituted by an A, A-, B+, B, B-, C+ and C scale to allow for more discrimination among students' performances.

Together with the revision of the rubric, the other forms used in recording oral assessments (oral assessment grade sheet and report card), were also modified. Furthermore, two additional instruments were developed: a feedback form used to record ongoing oral assessments; and a document titled "Guidelines for oral Assessment." This is a 12-page document which aims at offering teachers a theoretical and practical framework for assessing oral language and some guidelines that will foster the implementation of a homogeneous oral assessment system.

The revised Oral Assessment System consists of:

- Oral assessment rubric
- Redesigned Oral assessment grade sheet
- Redesigned Report card
- Feedback form
- Guidelines for Oral Assessment

In conclusion, since the concept of communicative competence is crucial to the assessment of spoken language, the aspects included in the rubric comprise linguistic competence (vocabulary, pronunciation, and grammar) as well as strategic, discourse and sociolinguistic competence (communicative effectiveness and task completion) and are described below.

- **Communicative Effectiveness:** ease with which students understand and deliver a message (smooth flow of speech). It also measures students' ability to use strategies to compensate for communication breakdowns and to initiate and maintain speech going. Features to keep in mind: Pausing/Hesitation (too long, unfilled pauses, chopped language); strategies such as circumlocution, self-correction, rephrasing, mimic, clarification, eliciting further information, comprehension checks, confirmation checks.

- **Grammar**: level of accuracy of previously studied structures. Students' grades should not be affected by lack of control of currently studied structures since such structures are not yet internalized. Features to keep in mind: form, word order, verb tense, subject-verb agreement, subject omission, etc.

- **Pronunciation**: level of correct pronunciation of already drilled sounds. Accent should not be penalized unless it interferes with communication. Features to keep in mind: Consonants, vowels, tone patterns, intonation patterns, rhythm patterns, stress patterns, and any other supra-segmental features that carry meaning.

- **Vocabulary**: extent to which the student uses vocabulary accurately, reflecting sufficient variety and appropriateness for the level and appropriateness to the context and interlocutor. Students should be able to incorporate vocabulary from previous levels. Features to keep in mind: rich vs. sparse, word choice, specific terminology, target-like phrasing.

- **Task Completion**: Accomplishment of the assigned task. A task is completed when students:

- Develop ideas with sufficient elaboration and detail (important information is not missing)

- Stick to the requirements (or steps) of the assigned task (in terms of functions of language: apologizing, requesting, inviting, etc)

The descriptors or criteria specified for the different aspects in the rubric allow teachers to interpret students' performance towards the accomplishment of the speaking standards established for each level at the Language Center.

The empirical investigation of validity was carried out by correlating the OAS against the speaking component of the Cambridge Examinations, KEY (Key English Test) and PET (Preliminary English Test). The results of this study are reported in Muñoz, el al. (2003b).

4. The Reliability of the Oral Assessment System

In oral assessment we are especially concerned with inter-rater reliability. Since we are aiming at a consensus on assessment, we need to decrease the unreliability due to disagreements in scoring among different evaluators. To evaluate the degree of inter-rater reliability, we conducted a pilot study where 14 randomly chosen students from different levels were evaluated by 5 raters during two stages: A and B.

Stage A

On Stage A, eight students were videotaped performing an oral assessment activity and were evaluated individually and scored by 5 raters using the modified rubric. For stage A, evaluators' mean scores were related through Spearman Rank Correlations for ordinal variables, and then compared using One Way ANOVA at a 0.05 level of significance. Spearman Correlations results are shown in Table 1.

This table shows Spearman correlations between each pair of variables. These correlation coefficients range between -1 and +1 and measure the strength of the association between the variables. Also shown in parentheses is the number of pairs of data values used to compute each coefficient. The third number in each location of the table is a P-value, which tests the statistical significance of the estimated correlations. P-values below 0.05 indicate the existence of correlations different from zero at the 95% confidence level. All pairs of evaluators show high correlations because p-values are close to 0.0.

However, Spearman correlations provide partial analysis because two variables may render high correlations but different mean scores. Thus it is necessary to conduct a One Way Anova analysis in order to have a more comprehensive picture of the evaluators' ratings.

The analysis revealed that two of the evaluators were in disagreement with the others. These results are given in Figure 1.

Evaluator 2 shows highly significant differences with respect to the other evaluators (p-value: 0.000). Likewise, evaluator 4 shows highly significant differences with evaluator 3 and moderately significant with 1 and 5. These findings indicated the need to either revise the rubric, discuss possible changes in some of its performance criteria or to train evaluators again to decrease differences and variability.

A revision of the rubric suggested that there were too many levels of descriptors (A, A+, B+, B, B-, C+, and C) which sometimes made it difficult for evaluators to discriminate among performance criteria. After discussing changes, the researchers decided to reduce the levels of descriptors and explain the others more precisely so that performance criteria could be easily differentiated. This would reduce misinterpretations on the part of the raters. It was also decided that letter grades would be replaced by numbers (1-5) on the grounds that:

- Teachers and students understand better numbers than letters
- Letters tend to carry emotional connotations that numbers lack, and
- Numbers allow for broader discrimination among performances.

Stage B

With the adjusted rubric, 6 students were evaluated individually on the same 7 aspects during stage B. Again, raters' scores were compared using Spearman Correlations and One Way ANOVA, results are shown in Table 2 and Figure 2.

Spearman correlations show a stronger association between the variables on stage B. One Way ANOVA showed no significant differences among evaluators (p-value > 0.05). This analysis reveals that inter rater studies need to be done regularly, not only during the research process, but also once the redesigned OAS is implemented. This is essential in order to maintain unified assessment criteria among raters. When inconsistencies are found, it will be necessary to determine which language aspect is causing discrepancy.

Table 1. Spearman Rank Correlations - Stage A

	EV1	EV2	EV3	EV4
EV1				
EV2	0.4695 (52) 0.0008			
EV3	0.6746 (52) 0.0000	0.6274 (52) 0.0000		
EV4	0.6368 (52) 0.0000	0.5736 (52) 0.0000	0.6387 (52) 0.0000	
EV5	0.6415 (52) 0.0000	0.6781 (52) 0.0000	0.6185 (52) 0.0000	0.4737 (52) 0.0007

Figure 1. One Way ANOVA - Stage A

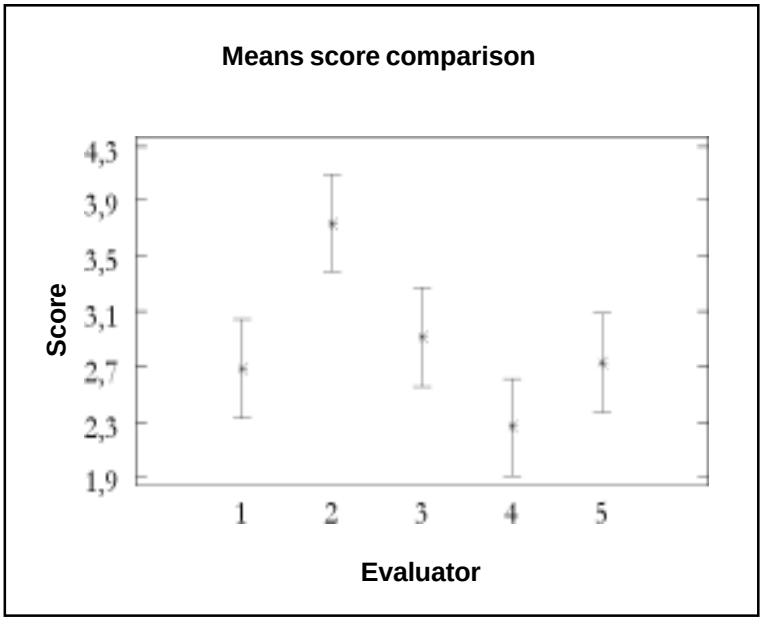
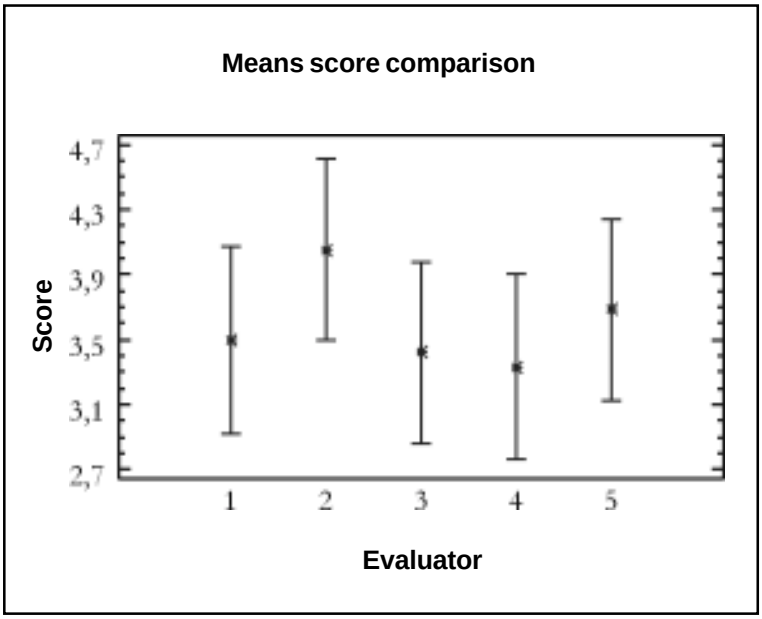


Table 2. Spearman Rank Correlations - Stage B

	EV1	EV2	EV3	EV4
EV1				
EV2	0.7256 (36) 0.0000			
EV3	0.7117 (36) 0.0000	0.8046 (36) 0.0000		
EV4	0.8448 (36) 0.0000	0.6500 (36) 0.0001	0.5493 (36) 0.0012	
EV5	0.8070 (36) 0.0000	0.7885 (36) 0.0000	0.7284 (36) 0.0000	0.7592 (36) 0.0000

Figure 2. One Way ANOVA - Stage B



5. Discussion

The results on inter rater reliability showed, on a first stage, significant differences among raters, although there was high correlation among them. Consequently, it was necessary to determine the cause of inconsistencies. Inconsistencies were found in:

1. The systematicity of evaluator 2 in assigning higher grades than the other evaluators
2. The variability regarding two aspects in the rubric: vocabulary and grammar

With respect to item one, and after discussions sessions were held, evaluator 2 was shown the need to standardize his scoring procedures. For item two, adjustments were made to the rubric, specifically, grammar and vocabulary were further detailed. It was thus possible to decrease differences among the evaluators. A second stage of evaluations showed that there were not significant inconsistencies. This clearly indicates that training and precise performance criteria are essential to increase reliability levels. With well-laid out criteria and well-informed raters, it is possible to reach an agreement in evaluation.

It is important to consider that this was a pilot study that helped on deciding on how to norm the evaluators and to establish the statistical tools to validate the results. Therefore, it is still necessary to conduct a study with a more representative sample.

Concerning validity, the logical analysis indicates that the OAS represents appropriately the construct communicative language competence as it is taken at the Language Center. Having a well-defined concept of the ability to be measured is important because if teachers have a good understanding of the ability they are measuring, their expectations of students' performance will be more realistic. Furthermore, they will be in a better position to make judgements or interpretations about the students' performance in a non-assessment situation.

The logical analysis also indicates that the rubric for oral assessment is a fairly valid instrument. The aspects included in the rubric are in agreement with the teaching methodology proposed by the Language Center and the theoretical definitions of communicative language competence. This means that the rubric represents adequately the construct communicative competence at the Language Center. Furthermore, the study on concurrent validity will provide further evidence to estimate validity.

6. Implications

First, we believe that training programs on oral assessment are necessary if we aim at a consensus on assessment. The redesigned OAS will be implemented in 2003. Before its implementation can be done, teachers need to be familiar with the system. To this end, an 8 – hour in-service training course - "Guidelines for Oral Assessment"- has been offered. The purpose of this course is to make sure all teachers from the Adult English Program are acquainted with the system. The objectives of the course are:

a. Ensure that teachers understand the conditions for effective assessment

b. Familiarize teachers with:

   - The definition of communicative competence
   - The revised OAS

A second implication of this study is the need to examine inter rater reliability frequently and systematically to ensure uniformity of assessment criteria.

Conclusions

The study on reliability provided a first look on consistency in the oral assessment instrument. We believe that these results are promising. They suggest that it is possible to obtain homogeneous scores among different raters in a short time of training. Furthermore, when there is a clear understanding of the ability that we want to measure, it is possible to create both assessment tools that reflect this ability and valid assessment tasks. As consequence, teachers will be able to evaluate their students more fairly and accurately, that is, they will be in a better position to make proper interpretations of their students' oral language ability.

Lastly, the development of assessment instruments that are valid and reliable is not an easy task. This implies a process of lengthy discussions and modifications that calls for support and encouragement on behalf of all the people involved in the educational system, mainly, the administration, teachers, and students.

Bibliography

Bachman, L. (1990). *Fundamental Consideration in Language Testing*. Oxford: Oxford University Press

Bachman, L. & A. Palmer. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Canale, Michael and Merrill Swain (1980). "Theoretical Bases of Communicative Approaches to Second language Teaching and Testing." *Applied Lingsuitics 1*.

Chomsky, Noam (1965). *Aspects of Theory of Syntax*. Cambridge, MA: M.I.T. Press.

Cohen, Andrew and Lydia Fass (2001). "Oral language Instruction: Teacher and Learner Beliefs and the Reality in EFL Classes at a Colombian University." Ikala Revista de Lenguaje y Cultura. Vol. 6, No. 11-12.

Flaitz, Jeffra (2000). *Communicative Language Teaching*. Unpublished Manuscript. Language Center, EAFIT University, Medellín, Colombia.

Genesee, Fred and John Upshur (1996). *Classroom Based-Evaluation in Second Language Education.* Cambridge: Cambridge University Press.

Herman, J.L., P.R. Aschbacher, & L. Winters (1992). A Practical Guide to Alternative Assessment. Alexandria, VA.: Association for Supervision and Curriculum Development.

Ingenkamp, K. and B. Wolf (1982). "Research in Oral Secondary School Leaving Examinations in Germany," British Journal of Educational Psychology 52, 341-349.

Jakobovits, L.A. (1970). Foreign Language Learning: a psycho-linguistic analysis of the issues. Newbury house.

Messick, S (1989). "Meaning and Values in test validation: the science and ethics of assessment." Educational Researcher 18 (2), 5-11.

Messick, S (1996). 'Validity and Washback in Language Testing.' Educational Testing Service.

Muñoz, Ana; Aristizábal, Luz D.; Crespo, Fernando; Gaviria, Sandra; Lopera, Luz A.; Palacio, Marcela. (2003a). Toward a Successful System of Assessment. Revista Universidad EAFIT , 129 January, February, March.

Muñoz, Ana; Álvarez Martha; Casals Sergi; Gaviria, Sandra y Palacio, Marcela (2003b) Validation of an Oral Assessment Tool for Classroom Use. Colombian Applied Linguistics Journal, 5, 139-157

Savignon, S.J. (1983). *Communicative Competence: Theory and Classroom Practice.* Reading, Mass.: Addison-Wesley.

Seward, B.H. (1973). "Measuring Oral Production in EFL." English Language Teaching Journal 28, 76-80.