ORAL TESTING

Laurie Frey de Charruí
Roger's Hall School
Mérida, Yucatán

Testing is an essential component of any ESL program. Oral testing is one type of measurement teachers use for evaluating language proficiency. Proficiency is defined as the ability to receive and transmit information within a real-life setting. In ESL teaching, oral testing is used to measure and evaluate the proficiency of non-native speakers.

Often oral testing includes both speaking and listening comprehension skills as contrasted to written testing which emphasizes those of reading and writing. For our purposes we will deal only with the testing of speaking skills.

When we test oral speaking we are really evaluating two major skills. One is linguistic competence. Linguistic competence includes accuracy of pronunciation, vocabulary, and structure. The other skill we measure is communication. Can the student speak about a practical topic with considerable ease? (Cohen 1980: 15).

In this article oral testing is discussed because I know many of us have problems deciding how to test orally. We will discuss the test taker, types of oral tests, scoring, and problems and considerations. Take what you already know about testing and incorporate these new ideas into each of your situations.

Before we consider the different types of exams we should think about our test takers. Students have different strengths and weaknesses with respect to language-learning abilities. The

following four points have an effect on how a particular student performs orally: his/her 1) verbal intelligence, 2) short term auditory and visual memory, 3) sound-symbol association skill, and 4) grammatical analysis. Thus, how the student speaks depends on skills acquired outside the classroom (Cohen: 25).

Personality also influences how the test taker performs. A study with Chicanos has linked speaking fluency with extroversion. In another study in Canada a four-trait factor was discovered: assertiveness, emotional stability, adventuresomeness, and conscientiousness all related to success in an oral exam (Cohen: 25).

We have seen that verbal intelligence, memory, association, analysis skills, and personality all play a part in our students' performance ability.

Oral exams come in a variety of formats and styles, and these differences may affect student output. Some students may do better with one exam than another. Fright does occur, but usually after the initial sessions it disappears. A good teacher can set the stage to minimize shock by recreating a friendly situation close to real life. On the other hand, the test is not a social session, and therefore the teacher should plan well (Jones 1975: 39).

Students need feedback on their speaking ability, and for that reason teachers should frequently check oral skills. There are various ways to evaluate your students. One is a grade based on progress in class. The student should be told at the beginning that he/she is going to receive a grade on class participation and that it is important. Then you would keep note-cards on each student and note each one's weaknesses and strengths. For example, if a sound is pronounced incorrectly repeatedly you would jot that down. The students could periodically have private conferences to discuss these points with you (Chastain 1976: 506-507).

Better than the in-class evaluation is a system of short speaking tests (Chastain: 507-508). There is no one design for these tests, and you are free to use a variety of techniques as you will soon see. If possible, the student should converse with a native speaker, or better yet, with a native other than the teacher. Test takers may sound more like natives if they are in dialogue with natives because the motivation is greater (Cohen: 123).

Keeping our test taker in mind, let's go on now to look at specific test formats. Examples of oral tests include real-life activities such as reading aloud, talking on the telephone or conversing face-to-face with the examiner or another person. Other exam formats are less realistic and use pictures or a reading as the stimuli for responses. The format that duplicates real-life situations is called direct testing. The second format is called indirect testing.

The following section illustrates the use of three formats with children. One example of an indirect test for children is story retelling. The teacher tells the child to listen closely to the story he/she is going to tell. Afterwards, another adult or teacher that has not heard the story comes in. The student retells the story and is graded for overall communicative effectiveness on a scale of one to five. The teacher also considers the following points: 1) whether to subtract points for insertion of extra material or to reward it as creativity; 2) how to weigh different points of the content and 3) how many points to subtract for omission of major details. Ideally, it is best if both teachers begin the session asking the student non-threatening questions about him/herself. By using a story about familiar objects, the teacher provides one good way of examining oral proficiency in children (Oller 1979: 367).

In the section on test formats for adult students we will see how cloze procedures may be used. This approach may also be

used with children, but the text and the instructions should be simplified. It is very helpful if examples can be given on the board before taking the exam (Oller: 367).

A third format to use with children is a direct method of testing. The child is given materials such as glue, scissors, yarn and paper and told to build something. Then he/she tells the teacher how the creation was made. Some teachers may like to give two grades, one for creativity and another for communication. Scoring can be based on the scales that will be mentioned later in this paper (Oller: 360).

Five years ago only a few methods were available for use with children. Now there are many more, in addition to all the measures for assessing adult proficiency (Cohen: 31).

Testing formats for adults seem to be more abundant because of the new importance placed on speaking and understanding a second language. Many agencies such as the Peace Corps and the Foreign Service Institute have developed their own procedures and scales. Much research has been done on these and other exams for adults. The following examples for adults are subdivided into two categories: indirect testing and direct testing formats.

Five examples of indirect testing are: 1) story retelling, 2) description of events, 3) "Mark the Picture," 4) reading aloud, and 5) oral cloze.

The first, story retelling, follows the same procedure as that for children. This time, however, the student may be given a text in English to read and interpret. An oral rendition that is distorted or partially comprehensible would receive only partial credit (Oller: 360).

The second indirect format is a description of events. Normally, describing one picture is not a narrative task unless it is part of a system of events. A chronological set of pictures

may show a cause and effect relationship. The student invents a story about them.

The third indirect format was developed at the University of Michigan. I call it, "Mark the Picture." It is a highly structured test in which the teacher gives the student a set of four pictures which differ only in certain respects. The examiner receives an identical set and sits at a distance so not to see the picture the student chooses. The student tells the examiner which picture to mark by describing it. The score is determined by the number of the correct guess in an allotted period of time (Oller: 325).

The fourth example, a more controversial format, is called reading aloud. It is very subjective and only for students that are known to be good readers in their native language. Reading aloud is easier than speaking; nonetheless it is unlikely that a person would be able to speak fluently with a text if he/she cannot even speak fluently without one (Oller: 332).

Scoring may include: time it took to read the article, accuracy, and word-for-word pronunciation. Ceiling rates, the maximum scores, can be calculated by giving the test to a native speaker. This is under the assumption that native speakers perform better than non-native speakers (Oller: 332).

The last example of indirect testing methods for adults is the oral cloze procedure. This is one of the least used oral test formats. Basically,it involves the student filling in blanks of a text with the missing words. To make the test, the teacher should select a non-controversial, non-technical article that is for a lower level. This is because we want to avoid putting a heavy burden on the memory and attention of the speaker. The text should be unfamiliar and might consist of representative samples from various sources put together. For the oral cloze the text is converted to an oral form such as a tape recording. Next, the kind of deletion procedure to use is decided upon. One example uses

a minimum of 50 pauses: beginning from a word near the front (sometimes the first sentence is left intact), every nth word is deleted until there are a total of about 50 pauses. For example, if the text is 250 words long, 50 deletions are possible if every fifth word is deleted. Unless you are assessing performance of specific grammatical structures, this nth word procedure is used, where "n" is any number. Finally, the deleted text form is recorded with or without inclusion of correct answers.

The text is first presented to the student in its entirety one or more times before being presented with pauses. After each pause the student answers orally and the correct answer may be given or not given to him/her before continuing (Oller: 332).

Another procedure for deletion is to take a 100-word passage and test every word. In the first run every fifth word is deleted. The second time every fourth word is deleted, and so on.

Still another procedure is to delete whole phrases, which is similar to the guessing we do in real-life communication. This is not a test of listening comprehension but an oral task which elicits speech. In reality, all normal speaking involves some listening comprehension.

Scoring the cloze procedure depends on the method used. If the student is rated for the exact work, creativity is inhibited. If answers are weighted in degrees of appropriateness, the scoring is more complex and takes more time. For example, four points could mean the exact answer, three points for an acceptable substitute, etc. Probably the best scoring method is accepting as correct any word that is in context. Unfortunately, this presents a problem of subjectivity (Oller: 366).

We have seen five examples of indirect testing methods for adults. Story retelling is a method that uses two examiners and can also be used for children. The description of events also involves a story, but this time produced creatively. "Mark the Picture" involves describing a single picture. Reading aloud is

a questionable method, and can only be used with students that are good readers in their native language. Lastly, oral cloze testing is a structured exam where students replace pauses with missing words.

Four new formats will be presented as examples of direct testing methods. If you remember, these tests are closer to real life and involve another person as the stimulus for response.

The first direct method we will look at is the mock lecture. Here the student takes notes on the main points of a particular lecture. He/she is graded on whether or not he/she noted the main points of the lecture, and also is asked either to summarize orally or to answer questions about the presentation. The problem with this is that it may be difficult to relay technical or unfamiliar vocabulary. Therefore, the lecture should be short and about a familiar topic (Cohen: 36).

The second format, interaction in groups, does just what it says; it tests the interaction that goes on when groups of two or more students are brought together. The groups are each given a topic from class work and their talking may be recorded on a tape recorder for grading later. Ideally, it is good to have a teacher present to prompt the students if necessary (Chastain: 507).

The third direct method, role-playing, is one of the most popular testing formats. Students are given roles and must interact with fellow students as that new person. This method is very difficult because the student is participating in real conversation with another person (the teacher), and he/she must be a listener as well as a speaker. Here are some examples:

1. Students are given secret instructions on how to play a particular role: one student is told in secret that he is a plainclothes policeman and he has just encountered the student activist responsible for a riot on campus

last week. Another student is told he is a withdrawn, serious, student. He is told a joker is loose, impersonating a policeman.

2. Students are given a situation and asked to react or say what they would say in the same situation:

   Your neighbor's dog is on the porch barking at the crack of dawn for the third day in a row. You go next door to complain. Neighbor, opening door with a big smile: "Come in pal."

   Student responds.

3. You are in a restaurant. The plate your food is served on is dirty. What do you do and say? (Cohen: 72).

A variation of the three role-plays above may be asking a student to describe the events he/she witnesses. This could be a role-play acted out or one on video tape:

4. A student comes in pretending to be a robber. He is holding a revolver (banana). The teacher is frightened and gives him/her the money. The robber exits quickly with the money, eating the banana, and dropping the peel in the wastebasket on the way out (Jones: 31).

Another type of oral test is the oral interview. This is probably the most valid approach to oral testing. It is a direct method of getting someone to speak. Three examples follow: 1) a personalized interview, 2) the Foreign Service Institute Oral Interview and 3) the Ilyin Oral Interview.

A personalized interview usually begins with simple social formulae in English, such as comments on the weather or questions like, "Is this the first time you've taken an oral test?" For low

level learners the teacher can prepare a long list of personalized questions. Then the list is used to randomly select five to ten questions for each student. As the student answers the teacher grades. Students are given a few seconds to think, but answers are expected promptly. Good students can answer ten questions per minute. Slower students need more time. A variation of this is to randomly draw names for each question asked (Cohen: 72).

The interview should be different for higher level students. Professional topics, current events and detailed aspects of jobs are good ways to put their ability to the test, and at the same time grammatical points can be elicited for checking (Jones: 31).

The Foreign Service Institute has developed an oral exam for adults, one which is planned closely with the purposes and objectives of their service. Two interviewers, a handout describing levels of speaking proficiency and a weighting scale are needed. The problem is the scorers must be trained to differentiate and have a knowledge of the levels of speaking and the weighting scale. It is also a time-consuming method and therefore a little expensive (Oller: 326).

The Ilyin Oral Interview, developed by Donna Ilyin, is a typical approach used by language teachers today. To administer the exam one "picture" is needed. Such pictures measure approximately 8 inches by 11 inches and are made out of sturdy paper. Markers are used to depict the colorful scenes and the necessary information (Oller: 316).

Thus the Ilyin Oral Interview has a structured format consisting of three sections. We can invent a number of questions from the events depicted to evaluate oral language proficiency in our students.

The personalized oral interview, the Foreign Service Institute Interview and the Ilyin Oral Interview are all ways to approach

the oral interview in a real-life situation. The test can be adjusted to the proficiency level of each student and can be done fairly quickly (Oller and Perkins 1980: 92). Students should be carefully prepared and provided with specific topics to focus on when studying. Unfortunately, it is nearly impossible to ask all students questions of equal importance and value.

Most of us often take scoring for granted. We may be more concerned about writing tests and fail to score competently. All tests for beginners must test knowledge of the sound system and grammar and vocabulary. At higher levels grammar and vocabulary are emphasized alone.

Both direct and indirect tests are usually administered by two examiners. One person leads the conversation and the other listens and makes notes of the strengths and weaknesses of the examinee. When both examiners mutually determine the score during, or at the end of the session, the scoring is called simultaneous. Simultaneous scoring allows the examiners to consider relevant stimuli such as facial expressions and lip movements. Also, the testing sessions may be shortened or lengthened, if necessary, by the examiners.

In addition to simultaneous scoring, there is delayed scoring in which responses are recorded for evaluation at a later time. This is best done by use of a video recorder, but a tape recorder is also possible. One advantage of this type of scoring is that unimportant variables such as attractiveness and mannerisms do not interfere with the testing. Also, there is opportunity for a playback of the communication to resolve any points of doubt.

Simultaneous scoring could be more reliable than delayed scoring. There are four reasons for this belief: 1) relevant stimuli are available such as lip movements, gestures, expression; 2) the rater has the conversation more clearly in mind; 3) the interview can be shortened or lengthened if necessary; and 4) there are no technical problems involved (Jones: 15).

Scoring in general requires human participation because there is no machine that is capable of evaluating speech in a conversation. But some day we will probably have something available that can imitate isolated sounds or phrases. Already there are computers based on speech-recognition devices (Jones: 18).

Presently, there are many human-scored rating scales of oral testing. One such scale involves grading the following four skills: pronunciation, grammar usage, vocabulary and fluency. This must be coupled with the following criteria for scoring:

5 - expresses himself well with practically no errors
3 - communicates fairly well but with noticeable errors
1 - practically incomprehensible
0 - no response

These numbers can be converted into any point scale desired.

Not surprisingly, some experts criticize this method. They prefer to omit fluency because it is difficult to assess speed. A more important reason is that people differ in speech rate in their native languages. In fact, hesitations such as "er" or "uh" may be signs of searching. It may be valuable to teach students how to be disfluent so they sound more native-like.

> To be fluent in the right way, one has to know how to hesitate, how to be silent, how to self-correct, how to interrupt and how to complete one's expression. According to this definition of fluency, one must speak in a way that is expected by the linguistic community and that represents normal, acceptable and relaxed language behavior. Testing of this quality of speech is not possible by means of any instrumental method (Cohen: 122).

Also criticized is the grading of grammar at beginning stages of learning. One expert likes teachers to overlook this

area if it does not interfere with communication. Perhaps grammar can be added later. He also questions the idea of rating vocabulary if it does not interfere with the communication (Cohen: 123).

Now let's look at another type of rating scale. This one is less threatening and more instructional than the traditional set of scales based on grammar, pronunciation, etc. I call this Levenston's scale. This type of scale helps students get an idea of what it means to gain control of oral abilities.

Each of the above two examples can be used with beginning ESL students, but students from a course stressing conversation and functional language can be expected to rate higher (Cohen: 124). Studies have found that there is not a one-to-one correspondence between what is taught and what is learned. Students may learn partially or even incorrectly.

Levenston's rating suggestion is broken into two main divisions: form and content. Each of these is then divided into three subclassifications. Form consists of:

a) naturalness of discourse: whether the student selects words appropriately.

b) style of expression: uses acceptable style such as passive vs. active tense.

c) clarity of expression and comprehensibility: message is understood.

All are graded on a one to five basis with five being the best score: a) is graded one, unnatural, vs. five, natural; b) is one, foreign, vs. five, native; c) is one, unclear, vs. five, clear.

The content division is graded into three areas also:

a) suitability: a gut feeling for whether the content could be native-like.

b) accuracy of information: ability to convey ideas accurately.

c) amount of information related: whether the student supplies the appropriate amount of information. (Many ESL students say more than a native would.)

These three subdivisions are also graded on a one to five basis: a) is graded one, unsuitable, vs. five, suitable; b) is graded one, inaccurate, vs. five, accurate; c) is one, inappropriate(too little or too much), vs. five, appropriate (Cohen: 121).

Another example of a rating scale format is a simplified measure of form and content. Form is defined as the way of expressing the message. Content is concerned with the message in the given situation. The scale is based on three points where one is the highest score and is received if the student uses appropriate form and content. Two points are for a message that is appropriate in content but not in form. Three points are given if the message is inappropriate in content. Three examples of what one student said follow:

Would you be free for coffee?	1 point
Because this plate is dirty, . . .	2 points
Can you prove it? (inappropriate in its context)	3 points

Lastly, for those who are interested in scoring oral recitations, a special procedure is used. Here the student is graded for ability to keep to the text without inserting other words, omitting them or changing word order. A point is given for every word recited correctly and one-half point for a repetition. One point is added if there is self-correction. In this way teachers must pay close attention and will not be misled by a person who

has excellent pronunciation but very poor reading (Cohen: 35).

Something often questioned among ESL teachers is whether or not a non-native speaker can rate the proficiency of the students. One researcher suggests that naive speakers might be better than native speakers who understand sophisticated language.

Another researcher expresses opposite views saying naive teachers may put too much importance on pronunciation, accent and fluency, and too little on the weightier grammar and vocabulary (Oller and Perkins: 103).

We have seen how four scoring approaches can be used in the ESL classroom. The four-criteria system is often criticized because its criteria are dangerous to measure on such a small sample of speech. Levenston's scale is probably the most non-threatening and detailed. The three-point scale is very simplified and based only on form and content. Finally, the last method describes one way we can score oral recitation.

Whatever the rating scale we choose, we should probably consider whether some criteria are more important than others and grade them accordingly. More importantly, not all categories may be relevant to our needs. Also, we must remember the element of time. If we give too much time to complete the exams, weak students will not be differentiated from good students in test scores (Oller: 325).

Before concluding we should take a look at some problems and considerations of oral testing. Most people are shocked when they hear about oral testing for the first time because they think it is too subjective. But it has been shown to be valid on a very professional level. To apply these tests to our classrooms we need to have well-defined criteria for testing and well-trained teachers that know how to use the definitions.

All good oral tests must have three basics: reliability, validity and practicality. A test is reliable if a student receives

approximately the same score on the same exam taken two different times. The average of several short quizes is more reliable than one long oral exam at the end of the course. It is up to the teacher to make sure her scores are reliable. Some suggest two people grade students' abilities. This too may pose a problem if one tester is dominant or has more experience than the other. It would be bad if the dominant rater had the wrong score. To avoid this it has been suggested that rating scales use ranges of scores which can be averaged (Jones: 40).

Oral exams are also criticized for not being valid. A test is valid if it tests what it is supposed to test, in this case real-life communicable ability. The Northern Regional Educational Laboratory evaluated 24 language tests and found that there is a serious need for valid tests (Oller: 307).

Thirdly, oral tests are not practical according to many people for reasons of cost, manpower and time for administration and scoring. Such testing does take time, but so does good teaching.

One other problem often mentioned is the lack of trained teachers. Teachers must know how to use rating scales and be able to give adequate attention to the student tested. Sometimes teachers participate too much in the oral exam, and the students do not present an adequate sample of speech. Moreover, the teacher must not confuse intellectual ability with linguistic ability. Often teachers fail to push the proficient students to their limits.

Despite these shortcomings there is a great need for oral testing. If a person can speak but not write a language we feel he knows the language, but if he cannot speak the language he cannot fully know it.

Teachers must work hard to make reliable, valid and practical tests. Students should be told to throw themselves into

the role of a native speaker. Since student interest is high immediately after a test, teachers should give prompt feedback on their scores and what they mean (Cohen: 40).

If you are having trouble developing oral exams, look back over the types of oral exam formats and methods to score them. Try to elicit natural responses through direct testing. If you can, use simultaneous scoring. By taking all the problems and considerations into account we can construct a proficiency test that is an accurate measure of our students' language ability. Oral tests can encourage or discourage our learners. Students should know what to expect in their exams and be relaxed. Oral exams should support, not surprise.

References

Chastain, Kenneth. 1976. Developing Second Language Skills. Chicago: Rand McNally.

Cohen, Andrew D. 1980. Testing Language Ability in the Classroom. Rowley, Mass: Newbury House.

Jones, Randall L. 1975. Testing Language Proficiency. Arlington, Virginia: The Center for Applied Linguistics.

Oller, John. 1979. Language Tests at School. London: Longman, 1979.

Oller, John and Kyle Perkins. 1980. Research in Language Testing. Rowley, Mass: Newbury House.