THE INTER-RATER RELIABILITY OF TWO ALTERNATIVE ANALYTIC GRADING SYSTEMS FOR THE EVALUATION OF ORAL INTERVIEWS AT ANADOLU UNIVERSITY SCHOOL OF FOREIGN LANGUAGES

A THESIS PRESENTED BY

ECE SELVA KARSLI

TO THE INSTITUTE OF ECONOMICS AND SOCIAL SCIENCES

IN PARTIAL FULFILLMENT OF MASTER OF ARTS

IN TEACHING ENGLISH AS A FOREIGN LANGUAGE

BILKENT UNIVERSITY

JULY 2002

ABSTRACT

Title: The Inter-rater Reliability of Two Alternative Analytic
Grading Systems for the Evaluation of Oral Interviews at
Anadolu University School of Foreign Languages

Author: Ece Selva Karslı

Thesis Chairperson: Dr. Sarah Klinghammer
Bilkent University, MA TEFL Program

Committee Members: Dr. William E. Snyder
Bilkent University, MA TEFL Program

Dr. Martin Endley
Bilkent University, School of English Language

Dilek Hancıoğlu
METU

Of all language exams, the accurate testing of speaking is regarded as the most challenging to prepare, administer and score because it takes considerable time and effort to obtain reliable results (Madsen, 1983; O'Malley & Pierce, 1996). Since subjective types of tests (e.g. interview ratings) require the judgment of the raters, inconsistency in judgments, which may affect the rater reliability adversely, may occur.

This research study investigated the inter-rater reliability of two alternative speaking assessment criteria designed for Anadolu University, School of Foreign Languages. The perspectives of the participants on the scales were also analyzed with the help of the interview records.

Two types of data were used in this study: raters' scores using both of the scales and raters' opinions of the rating scales. The participants in the study were five English instructors currently employed at Anadolu University School of Foreign Languages.

The teachers attended the training and norming sessions for the four-band scale and then graded 36 elementary level students' oral performance using the scale. Then the teachers were interviewed as a group. They were asked to express their opinions about the scale. Six weeks later, same procedure was followed for the five-band scale. The training and norming sessions for both of the scales were held by the researcher.

Then inter-class correlation for both of the scales was calculated using the scores assigned to 36 elementary level students. The result of the statistical analysis revealed that the four-band scale is more reliable than the five-band scale.

The results of the interviews indicated that the raters have common problems in assigning the scores to students' oral performances while using both of the scales. The problem that the raters faced in the scoring procedure while they were using the five-band scale is that two terms used in the descriptors are not clear. The common problems faced by the raters while they were using the four-band scale are as follows: 1) one term used in the descriptors is not clear, 2) students' performance may not fit into the bands, 3) the number of bands in each category is not enough, and the highest band in vocabulary needs to be more detailed 4) the lowest band is unnecessary, 5) there is a big difference among the bands in terms of the value assigned to each band.

After an analysis of the two speaking assessment scales, the four-band scale is recommended to assess oral performances of elementary level students' at Anadolu University School of Foreign Languages. Since nearly all participants stated problems concerning the descriptors in both of the scales, the descriptors need to be reconsidered and paid more attention to during training and norming sessions. In addition, the scale is open to revision in terms of weighing because the participants had problems with it. Finally, it is recommended that teachers who are going to take part in the assessment of learners’ oral performances need to attend training and norming sessions before they take part in the actual scoring procedure.

BILKENT UNIVERSITY

INSTITUTE OF ECONOMICS AND SOCIAL SCIENCES

MA THESIS EXAMINATION RESULT FORM

JULY 3, 2002

The examining committee appointed by the Institute of Economics and Social Sciences for the thesis examination of the MA TEFL student

Ece Selva Karslı

has read the thesis of the student.

The committee has decided that the thesis of the student is satisfactory.

Thesis Title:	The Inter-rater Reliability of the Two Alternative Analytic Grading Systems for the Evaluation of Oral Interviews at Anadolu University School of Foreign Languages
Thesis Advisor:	Dr. William E. Snyder Bilkent University, MA TEFL Program
Committee Members:	Dr. Sarah Klinghammer Bilkent University, MA TEFL Program
	Dr. Martin Endley Bilkent University, School of English Language
	Dilek Hancıoğlu METU

We certify that we have read this thesis and that in our combined opinion it is fully adequate, in scope and quality, as thesis for the degree of Master of Arts.

Dr. Sarah Klinghammer
(Chair)

Dr. Martin Endley
(Committee Member)

Dilek Hancıoğlu
(Committee Member)

Dr. William E. Snyder
(Committee Member)

Approved for the
Institute of Economics and Social Sciences

Kürşat Aydoğan
Director
Institute of Economics and Social Sciences

ACKNOWLEDGEMENTS

I would like to express my special thanks to my thesis advisor, Dr. William E. Snyder for his invaluable guidance, constant encouragement at every stage of this thesis study.

I am thankful to committee members Dr. Sarah Klinghammer, Dr. Martin Endley, and Dilek Hancıoğlu who enabled me to benefit from their expertise.

I am deeply grateful to the Director of the Preparatory School of Anadolu University, Prof. Dr. Gül Durmuşoğlu Köse, who provided me the opportunity to study at MA TEFL Program at Bilkent University.

Thanks are extended to Prof. Dr. Hüsnü Enginarlar, Prof Dr. Ayşe Akyel, Dr. Sarah Klinghammer, Julie Mathews Aydınlı, Hossein Nassaji, Doç. Dr. Handan Kopkallı Yavuz, Gülsüm Müge Kanatlar, and Dr. Şeref Hoşgör who provided me with invaluable feedback and recommendations.

I owe the greatest gratitude to students of elementary 10 and 11 classes, Hülya İpek, Gaye Çalış Şenbağ, Tuba Yürür, Meral Melek Ünver, and Dilek Altundaş who participated in this study. Without them, this thesis would never have been possible.

I am sincerely grateful to all my MA TEFL friends, especially Emel Şentuna and Aliye E. Kasapoğlu, for their cooperation, support, patience, and friendship throughout the program.

Finally, I must express my deep appreciation to my dear fiancée and my family, who have always been with me and supported me throughout.

To my present and future families

for their endless support and love…

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Background of the study

Performance assessment has become increasingly popular in the language teaching field to test communicative competence since the focus in the language classroom is on communicative language teaching in recent years. Second language oral testing increasingly calls for more performance-based tests (Chalhoub-Deville, 1996, McNamara, 1996).

Performance-based assessment requires a candidate to use language in some way while a judge evaluates the performance (McNamara1996). Gronlund (1998) points to a number of advantages of performance-based assessment over traditional assessment. Performance-based assessment allows direct evaluation of what learners can do with the language rather than what they know about it, as traditional tests. It provides greater motivation for students by making learning more meaningful by providing more authentic testing of what has been studied. However, performance-based tests also have some limitations. In particular, the scoring is subjective and may have low reliability.

Today many institutions are testing students' competence in speaking through performance-based tests such as, interviews and oral presentations, because good classroom testing is related to what has been taught (Hamp-Lyons, 1990, Hughes, 1989). If the communicative language teaching approach is used in the language classes, performance-based assessment needs to be used.

The accurate testing of speaking is widely regarded as challenging because it takes considerable time and effort to obtain reliable results (Madsen, 1983, O'Malley & Pierce, 1996). One reason is that speaking has many components (e.g., fluency, and accuracy) and it is difficult to define them. Because the components of speaking ability cannot be identified easily, what criteria to choose in evaluating oral communication and how to test and weight them are problematic (Madsen, 1983). When there are a large number of test takers, practical constraints on time and other resources may affect the quality of testing. It may not be possible to train and norm examiners adequately (Cohen, 1980, Hughes 1989, Weir, 1990, 1995). Most important, the subjective nature of the scoring procedures involving human judges can affect the scorer reliability negatively (Brown, 1996, Harris, 1969) Because performances are not usually recorded and cannot be checked later, creating an assessment system that minimizes these potential negative effects on reliability is essential (Weir, 1990, 1995).

Brown (1996) defines test reliability as "… the extent to which results can be considered consistent" (p. 192). One type of test reliability is rater reliability. Some authors (Brown, 1996, Hughes, 1989, Lado, 1961) have suggested how high a reliability coefficient we should expect from oral production tests. The reliability coefficient of oral production tests should be in the .70 to .79 ranges, which is considered adequate for oral tests. Since raters are necessary when testing students' productive skills through performance tests, testers most often rely on rater reliabilities as a measure of test reliability in such situations (Brown, 1996).

It is possible to minimize the effect of these factors and maximize reliability if a rating scale is designed with a clear and concise description of performance at each level (Bachman, 1990, Heaton, 1994). In addition, reliability can be increased by using more than one assessor (Bachman, 1990, Brown, 1996, Underhill, 1987, Weir, 1990, 1995). Training and norming sessions are also crucial in obtaining reliable scores. During training and norming sessions, the raters become familiar with the rating scale and learn how to apply it consistently (Alderson, Clapham, & Wall, 1995). Given these conditions, the inter-rater reliability of rating scales used can be analyzed in order to find out whether the scales are adequately reliable for the institution.

Context of the Study

Anadolu University School of Foreign Languages was established in 1998 and served nearly 2000 students during the 2001-2002 year. The number of the instructors currently employed at the institution is 82.

Students who are not proficient in English are required to study at the university preparatory school for one year. At the beginning of each term, students are placed in appropriate levels according to their scores in the placement exam. The levels are beginner, elementary, low-intermediate, intermediate, upper-intermediate and advanced.

As the program is skills-based, each skill is taught separately and assessed separately. In beginner, elementary and low-intermediate classes, four hours a week are devoted for speaking, while in other levels two hours a week are spent on it. Speaking is assessed three a year. In each semester, there is one mid-term exam, which is held as an achievement test. Each speaking exam comprises 20% of the total score in a term. At the end of the year, the students in all levels are required to take the final test. The final test has three sections, one of which is a speaking exam. The speaking section comprises one third of the total score of the final exam. Speaking tests are important because the results of these tests help determine whether the students can pass the preparatory class and attend their own faculties.

All instructors take part in speaking assessment, even if they have not been teaching speaking during the term. Two instructors, both in midterms and the final exam, assess students in pairs through interviews. The instructors use a speaking assessment scale. There are three categories in the scale currently used: task achievement, fluency, and accuracy and appropriacy (see Appendix A). Each category is weighted equally. Raters assess each category out of hundred and take the average score as their final result. Coordinators calculate the average of the raters' scores as the final grade. Although a standard form of a rating instrument is used and two teachers assess the same learner, there are still sometimes inconsistencies between teachers.

A number of factors underlie these inconsistencies. The criterion itself has design problems with both its categories and the bands within them. For example, one of the categories 'accuracy and appropriacy', is too broad because it assesses appropriate use of not only grammar but also vocabulary in a category. Learners may not perform equally in terms of grammar and vocabulary and the performance may not fit into a common band (Hughes, 1989). For each category, the top band can be scored as 100, 95, 90, or 85, but there is one description of performance for these four possible grades. Raters are not provided with descriptors that differentiate the scores in each band; therefore, they may assign the scores inconsistently. Lack of teacher training with the assessment criterion may be another reason. It is hard to conduct training and norming sessions at Anadolu University School of Foreign Languages because of the large number of teachers involved in oral assessment procedures. Their heavy workloads and differing schedules preclude the organization of training sessions.

In addition, resources to conduct training and norming sessions for assessment of oral interviews are limited. There are not any video recordings of sample student interviews to use in the session. In each mid-term exam, different test tasks are chosen for a level according to the syllabus. Also, different test tasks in different levels are used in the interviews. Therefore, video recordings of student interviews from different levels are needed to use in the sessions. In brief, teachers cannot be standardized in using the rating scale as no training and norming sessions held at Anadolu University School of Foreign Languages. In conclusion, a scale that produces reliable scores with minimum amount of training and norming sessions is needed for the sake of practicality.

Statement of the Problem

As tests play an important role in making decisions about students' performance and level of knowledge, they need to be scored consistently. Because of the nature of performance tests (e.g., interview ratings in speaking), it is difficult to obtain reliable scores. These tests require the subjective judgment of the raters (Brown, 1996, Harris, 1969). Brown (1996) puts the problem as "… the subjective nature of the scoring procedures can lead to evaluator inconsistencies or biases having an affect on students' scores and affect the scorer reliability adversely" (p. 191).

The use of a well-designed rating scale and multiple, trained raters helps increase the reliability of performance assessment and makes the assessment process one that gives meaningful results (Alderson, 1995, Underhill, 1987).

Although a rating scale is used and two teachers assess the same learner at Anadolu University School of Foreign Languages, there are still sometimes inconsistencies between teachers. The organization of the current scale was judged inadequate by the administration of Anadolu University School of Foreign Languages. A decision was taken to change the current scale to improve the scoring of speaking tests. I was asked to design a new scoring criterion for this purpose. I produced two and will compare their inter-rater reliability here. In order to help increase reliability, minimal training sessions were included in my design. The goal is to find the criterion which is most practically reliable.

Purpose of the Study

The purpose of this research study is to investigate the inter-rater reliability for two alternative oral assessment scales designed by the researcher for Anadolu University School of Foreign Languages. Teachers' perspectives on the use of the two alternative speaking assessment scales will also be examined.

If the results of the study show that one and/ or both of the alternative speaking assessment criteria can be considered adequately reliable in terms of inter-rater reliability, the researcher will make some recommendations for the two criteria and propose a suggested speaking criterion for Anadolu University School of Foreign Languages.

Research Questions

This study will address the following research questions regarding speaking assessment at Anadolu University:

1. What is the inter-rater reliability of the four-band speaking assessment scale developed to be used at Anadolu University School of Foreign Languages?

2. What are the participants’ perspectives on the use of the four-band speaking assessment scale?

3. What is the inter-rater reliability of the five-band speaking assessment scale developed to be used at Anadolu University School of Foreign Languages?

4. What are the participants’ perspectives on the use of the five-band speaking assessment scale?

Significance of the study

Two speaking achievement exams are given at Anadolu University School of Foreign Languages. Since students are required to take the speaking exam and the results of the exam play an important role in making a decision about students' performances, a reliable speaking assessment criterion is needed.

The use of a reliable assessment instrument will help instructors to test more accurately and comfortably because inconsistencies between raters may be reduced. Learners will receive more accurate marks and may feel more positive about the assessment procedure as a result.

All administrators and EFL teachers who have difficulties in assessing learners’ speaking performance may benefit from this study. This research study will also be valuable for other people in other institutions who would like to use an analytic grading system to score learners’ oral performances. They may take this research study as a model and investigate the inter-rater reliability of their own rating scales.

CHAPTER 2: REVIEW OF LITERATURE

Introduction

This research study investigated the inter-rater reliability of two alternative speaking assessment criteria designed for Anadolu University School of Foreign Languages. In addition, the perspectives of the participant raters on the two alternative criteria will be analyzed with the help of interview recordings done after workshops employing each scale. Based on the results gathered from the statistical and interview analysis, recommendations will be made about the use of the two alternative scoring systems.

This chapter reviews the literature on testing speaking. The chapter consists of five sections. In the first section, the literature on performance-based assessment will be briefly reviewed, including information on its strengths and limitations, formats of testing speaking, and problems of testing speaking. The second section covers reliability in relation to the scoring of students' oral performance. The third section examines the rating scales, including advantages and disadvantages of analytic scales. The fourth section looks at designing criteria for oral performance tests and problems in developing criteria. Finally, the fifth section discusses the importance of training raters in scoring oral performance.

Performance-based Assessment

Many experts (Brown, 1996, Chalhoub-Deville, 1996, McNamara, 1996, O'Malley & Pierce, 1996, Underhill, N. 1987, Weir, 1990, 1995) state that since the emphasis in the language classroom began to move from the classical approaches in instruction and testing to a more communicative approach, classroom teachers and researchers have had to address the problem of how to measure students' performance. Communicative teaching techniques and styles present a particular problem. These techniques and styles aim to change the traditional language learning approach and that implies that the method of evaluation must also change. Savignon (1983: 246) pinpoints the problem of evaluating communicative competence and states, “The most important implication of the concept of communicative competence is undoubtedly the need for tests that measure an ability to use the language effectively to attain communicative goals” (cited in Edelman, 1987). Second language oral testing increasingly calls for more performance-based tests.

McNamara (1996) distinguishes the format of a performance-based assessment from the traditional assessment by the presence of two factors: “… a *performance* by the candidate which is observed and judged using an agreed *judging process*” (p. 10). In addition, these tests often employ more than one test method. Consequently, the test method and the rater become integral components of performance-based tests, influencing test scores (Chalhoub-Deville, 1996).

Gronlund (1998) presents a variety of strengths of performance assessments. They permit the evaluation of skills that cannot be tested in traditional ways, allowing testers to see whether students can use their knowledge in action. In addition, performance assessment provides a “more natural, direct, and complete evaluation of some types of reasoning, oral and physical skills” (p. 137). By basing test tasks on real world problems and situations performance assessments help motivate students and provide them with clear goals for learning. The result makes the learning process more meaningful.

In addition to all these advantages, Gronlund (1998) mentions some practical limitations of performance assessment as well. They require considerable time and effort to use. Evaluation must frequently be done individually, rather than in groups. Having these individuals perform enough tasks to be able to judge their abilities requires extra time. Judging and scoring learners' performances is subjective and may have low reliability. Using human judges creates inherent inconsistencies in the process which needs to be controlled.

Students' oral ability is usually assessed through use of performance assessments. In recent years, oral performance is assessed in many schools and institutions all over the world (Cohen, 1994, Gronlund, 1998, McNamara, 1996, Weir, 1990, 1995). The formats of testing speaking can be grouped under two headings: direct tests, including interviews, role-plays, and indirect tests, such as prepared monologue, and reading aloud (Carroll and Hall, 1985, Harris, 1969, Hughes, 1989, Weir, 1990). These formats will be explained in detail below.

Formats of Speaking Tests

Hughes (1989) lists three common formats for speaking tests: interviews, interaction with peers, and response to tape recordings. The three formats and their advantages, and disadvantages are as follows:

Interviews are the most common format for testing speaking. There are two types of oral interviews: the free interview and the controlled interview. In the free interview, no set of procedures for eliciting the language is laid down in advance. Since differences may occur in the interviews, the performances are likely to differ from topic to topic that the learners are supposed to speak on. As bands in the scale include a limited number of descriptors, matching each performance with the scale becomes more difficult. Also, the procedure is time consuming and difficult to administer if there are a large numbers of candidates (Cohen, 1980, Harris, 1969, Weir, 1990).

In the controlled interview, a set of procedures is determined in advance for eliciting performance. The controlled interview has some advantages. First, since the candidates are asked the same questions, it is easier to compare the performances. Second, it has been shown that with sufficient training and standardization of examiners to the procedures and scales employed, reasonable reliability figures can be reached. Clark and Swinton report average intra-rater reliabilities of 0.867 and inter-rater reliability at 0.75 for FSI type interviews, which is close to the model of controlled interviews (cited in Weir, 1990). One of the drawbacks of the controlled interview is that it cannot cover the range of situations which the candidate might have to perform in in real life. Besides that, there is still no guarantee that the candidates will be asked same questions in the same manner, even by the same examiner (Weir, 1990).

Weir (1990) states that the common advantage of oral interviews is that they have a high degree of content and face validity. Therefore, they are a popular means of testing the speaking skills of candidates. The most frequently employed method in scored interviews is to have one or two trained raters interview students either individually or in very small groups and record the performance. If the interview is recorded, raters can have a chance either to score or check the performance later.

If interviews are not designed appropriately, they may have one serious drawback. Hughes (1989) states, “The relationship between the tester and the candidate is usually such that the candidate speaks to a superior and is unwilling to take the initiative” (p. 104). As a result, only one type of speech is elicited, and many functions are not represented in the candidate's performance. In order to overcome this problem, a variety of techniques need to be used during the interviews.

Interaction tasks are another common format for speaking tests. There are two types of interaction tasks: student-student information gap and student-examiner information gap (Hughes, 1989, Weir, 1990). These types are discussed in detail below.

In student-student information gap two or more candidates are given a task. They may be asked to discuss a topic or make plans. The main advantage to this format is that the task is highly interactive since the students must use question forms, ask for clarification, and elicit information in order to complete the task. Therefore, "the task is highly interactive and as such comes much closer than most other tasks to representing real communication" (Hughes, 1989 p. 78). The problem is that the performance of one candidate is likely to be affected by that of the other. Similarly, if there is a big difference in proficiency between the two students, this may influence performance and also the judgment made on it. It is suggested that the candidates need to be either free to choose their partners or carefully matched if this format is used.

The second format for interaction tasks is student-examiner information gap. In this format, students separately can be given a set of notes or diagram that has some missing elements and their task is to request the missing information from the examiner. In general, a common interlocutor, for example, a familiar teacher with whom the students would feel comfortable is employed to conduct the test. Weir (1990) states the main advantage as "There is a stronger chance that the interlocutor will react in a similar manner with all candidates allowing a more equitable comparison of their performance" (p. 79). The disadvantage is that interacting with a teacher is often "a more daunting task for the candidate than interacting with his peers" (p. 179). During the test students may feel that they are not equal in status although a friendly and familiar teacher is generally chosen. This may affect students' performances negatively.

Response to tape recordings is the third format for speaking tests (Hughes, 1989). All candidates are presented tape-recorded stimuli only with the same audio or video. The advantage of this format is large numbers of candidates can be tested at the same time if a language laboratory is available. One problem with this type of speaking test is the use of audio or visual aids might be stressful to some candidates. Another disadvantage of this format is that there is no way of following up candidates' responses.

In addition to these formats, Weir (1990) adds two more ways of conducting speaking tests, which are verbal essay and oral presentation. In verbal essay, students are asked to speak for three minutes on either one or more specified general topics. They are sometimes asked to speak directly into a tape recorder. One problem with this type of speaking test is about the choice of the topic. If open-ended topics are chosen, students may need more background or cultural information to be able to complete the task adequately. Therefore, it may be difficult to compare learners' performances and assess them consistently.

In oral presentation, the student is expected to give a short talk on a topic. He may be asked to prepare his talk beforehand or be informed about the topic shortly before the test. The advantage of this test is that the task is closer to real life tasks that the candidate might perform in the target situation, if the activity is integrated with previously studied texts. There is a danger that the student may learn the speech by heart. If little time is given for preparation to avoid students memorizing their talks, then there is a problem of what to test: topical knowledge or language ability. For example, although the students speak well, they may not give adequate information about the topic as they may not have background knowledge about it. Or, the candidate may know the topic well but cannot express this because of inadequate or limited language ability.

Problems of Testing Speaking

Madsen (1983) mentions a number of reasons to why speaking tests seem so challenging. The nature of the speaking skill itself is not usually well defined; and therefore there is some disagreement on just what criteria to choose in evaluating oral communication. Grammar, vocabulary, and pronunciation are often measured and named as aspects of speaking skill. Other factors such as fluency and appropriateness are also usually considered. But there are still other factors such as listening comprehension, correct tone (e.g., sadness or fear), reasoning ability, asking for clarification to be identified in oral communication. Moreover, even when there is agreement on which factors to test in oral communication, there can be questions about how to test and weight each factor. Briefly, the elements of speaking ability are numerous and not always easy to identify or assign appropriate values to.

There are also practical constraints on testing spoken language proficiency. These include the administrative costs and difficulties of testing a large number of students either individually or in very small groups. Resources necessary for training and standardizing the examiners, paying a large number of examiners, and total amount of time needed for administering the speaking tests may not be sufficiently available (Hughes 1989, Cohen, 1980, Weir, 1990, 1995). Weir (1990) illustrates this situation and claims that most GCE Examining Boards in England were said to lose money on every candidate who sits an "O" level language examination in which there is an oral component.

In addition to these problems, the number of people involved in the interaction in the test is an important point (Underhill, 1987, Weir, 1995). Having one rater or two raters affects the scores assigned to students. If two raters are present in the test, the scores are combined. If one rater is present in the test, his score is assigned to the student. The reliability of the scores can also be affected. Underhill (1987) states "The more assessors you have for any single test, …. the more reliable that score will be" (p. 89). It needs to be considered when conducting speaking tests that the number of the raters in scoring is a factor in speaking tests. In addition, the role of the examiner and the interlocutor need to be identified well. Weir (1995, p. 41) indicates "If the examiner is also an interlocutor then the problems are further compounded". It becomes a harder to assign scores to learners if an interlocutor is a rater at the same time.

Assessing oral ability reliably is considered even more problematic because the performance is usually not recorded. This may cause problems because scoring takes place either while the performance is being elicited or shortly afterwards. Raters need to follow the interview and score the performance at the same time or shortly afterwards. In addition, if interview is not recorded, the performance cannot be checked later. Therefore, raters have to score the performance during or just after the interview (Weir, 1990, 1995).

Alderson, Clapham & Wall (1995) claim that one of the characteristics of the scoring of oral ability is that it is generally highly subjective. Oral tests are usually human-scored, meaning that raters assign scores. Examiners are required to make judgments about students' oral performances. Therefore, human errors in doing the scoring are another common source of measurement error (Brown, 1996, Harris, 1969). Chalhoub-Deville (1996) mentions the influence of the rater on scores obtained as a potential source of error that may influence learners' scores in second language oral ability. Brown (1996) states the problem as follows: "… the subjective nature of the scoring procedures can lead to evaluator inconsistencies or shifts having an affect on students' scores and affect the scorer reliability adversely" (p. 191). He illustrates the situation as follows:

> For instance, if a rater is affected positively or negatively by the sex, race, age or personality of the interviewee, these biases can contribute to measurement error. … Perhaps one composition rater is simply tougher than the others. Then a student's score is affected by whether or not the rating is done by this particular rater (p. 191).

Since any of the more subjective types of tests (e.g. interview ratings) requires the judgment of raters, minimizing these inconsistencies is an important part of ensuring fair scoring.

Reliability

Bachman and Palmer (1996) claim that the most important quality of a test is its "usefulness" and define usefulness as " … a function of several different qualities, all of which contribute in unique but interrelated ways to the overall usefulness of a given test" (p. 18). Reliability, construct validity, authenticity, interactiveness, impact, and practicality are the six qualities mentioned in the notion of usefulness.

Test developers need to find an appropriate balance among these qualities according to their purpose, students, and situations. Therefore, minimum acceptable levels for each quality will vary from one testing situation to another. In order to increase the reliability to a minimum accepted level in a testing situation, resources available in that context are important.

Reliability is defined as the extent to which results can be considered consistent (Bachman & Palmer, 1996, Brown, 1996). Alderson, Clapham, & Wall (1995) highlight the importance of validity and reliability in testing and state that if the marking of a test is not valid and reliable then all of the other work undertaken earlier to construct a “quality” instrument will have been a waste of time. Reliability is important in oral tests and studies investigating the reliability of oral interviews have conducted (e.g., Engelskirchen, Cottrell & Oller, 1981, Jones, 1979, Shohamy, 1981).

Reliable test scores are desirable because language teachers and administrators do not want to base their decisions about students’ performance on test scores that are inconsistent. These decisions are important decisions and can make big differences in the lives of students. As teachers and administrators are responsible in making such decisions, they want to have as accurate and consistent scores as possible (Brown, 1996). Getting scores from a test is a three-step process. First, the construct to be tested must be defined. Then, how the construct will be tested must be determined. Finally, a scoring method must be designed (Bachman & Palmer, 1996, Brown, 1996, Cohen, 1994, Harris, 1969, Hughes, 1989, McNamara, 1996, 2000).

For oral performance tests, the scoring method involves raters using a scale. Rater reliability is one type of reliability. Rater reliability is divided into two categories: 'intra-rater reliability' and 'inter-rater reliability' (Alderson, Clapham, & Wall, 1995). An examiner is judged to have intra-rater reliability if she or he gives the same marks on two different occasions. An inconsistent examiner is the one who changes his her standards during marking and or who applies the criterion inconsistently (Alderson, Clapham, & Wall, 1995). Since the focus of this study is on inter-rater reliability, it will be discussed in more detail below.

Inter-rater reliability refers to "the degree of similarity between different examiners" (Alderson, Clapham, & Wall, 1995, p 129). Two markers may differ enormously in respect to spread of marks and expectations. Heaton (1994) illustrates this situation in the following example.

> Marker A may give a wider range of marks than marker B, marker C may have much higher expectations than marker A and thus mark much more strictly awarding lower marks to all the compositions, and finally marker D may place the compositions in a different order of merit (p. 144).

It is not possible for all examiners to match one another all the time. However, it should be possible for raters to achieve adequate levels of consistency (a correlation coefficient of 070 or above; see Brown, 1996, Hughes, 1989, Lado, 1961, McNamara, 2000). This can be achieved through the use of a clear and practical rating scale and adequate training of raters (Alderson, Clapham, & Wall, 1995).

In addition, according to Underhill (1987) the most effective way of increasing reliability is to use more than one assessor. He also states "… two assessors, whose marks are combined, produce a more reliable score than a single assessor" (p. 90). One solution to reliability problem is to have more than one assessor for the test.

It is possible to minimize the effect of rater inconsistencies and maximize reliability if a rating scale is designed that describes performance clearly across all levels. Rating scales play a key role in increasing the reliability as they encourage raters to be consistent in their grading. A carefully designed rating scale enables the rater to identify what he or she expects for each band and assign the most appropriate grade to a student's performance being assessed. It also encourages raters to be consistent in their grading (Bachman, 1990, Heaton, 1994).

Rating Scales

Rating scales help increase the reliability of performance assessment and provide a common standard and meaning for the rating process (Alderson, 1995). Also, Stiggins (1987) stresses the importance of the statement of performance criteria as follows:

> No other single specification will contribute more to the quality of your performance assessment than this one. Before the assessment is conducted, you must state the performance criteria, in other words, the dimensions of examinee performance (observable behaviors or attributes of products) you will consider in rating…. Performance criteria should reflect those important skills that are the focus of instruction. Definitions spell out what we, as the evaluators, mean by each criterion (p. 20, cited in McNamara, 1996).

Gronlund (1998) defines the rating scale as follows:

> The rating scale is similar to the checklist and serves somewhat the same purpose in judging procedures and products. The main difference is that the rating scale provides an opportunity to mark the degree to which an element is present instead of using the simple "present-absent" judgment" (p. 154).

Murphy (1979, p.19) explains the nature of the marking scheme as "… a comprehensive document indicating the explicit criteria against which candidates' answers will be judged: it enables the examiner to relate particular marks to answers of specified quality" (cited in Weir, 1990).

Underhill (1987) agrees the definitions stated above and states the following:

> "A rating scale is a series of short descriptions of different levels of language ability. Its purpose is to describe briefly what the typical learner at each level can do, so it is easier for the assessor to decide what level or score to give each learner in a test. The rating scale therefore offers the assessors a series of prepared descriptions and she then picks the one which best fits each learner" (p. 98).

Rating scales are significant in certain types of performance assessment, as they are used to guide the rating process. Certain features of performance are determined and agreed. This involves various components of competence, such as fluency, accuracy, and sociocultural appropriateness. The weighing of each of the components is another important issue in performance assessment. (McNamara, 2000)

Different scales focus on different aspects of language use and for this reason different criteria are used for describing levels (Bachman & Cohen, 1998, Bachman & Palmer, 1996, McNamara, 1996). There are two different scoring systems used in assessment criteria: holistic and analytic scoring. Since analytic scales are used in this study, they will be discussed in more detail below.

Analytic Rating Scales

Analytic scoring is defined as a method of scoring which requires a separate score for each of a number of aspects of a task. It calls for the use of separate scales, each assessing a different aspect of performance such as grammar, vocabulary, and appropriateness. Each component is scored separately and sometimes given different weights to reflect their importance in instruction. A student's total score is the sum of the component scores (Alderson, Clapham, & Wall, 1995, Bailey, 1998, Cohen, 1994, Hamp-Lyons, 1990, Heaton, 1990, Hughes, 1989, Weir, 1995).

There are a number of advantages and disadvantages to analytic scoring which are explained below.

Advantages

There are a series of advantages to analytic scoring. As Hughes (1989) mentions, "Analytic scoring disposes of the problem of uneven development of subskills in individuals" (p. 94). Since learners are in the process of mastering the language, they may perform well in terms of one aspect of performance (e.g., fluency) but may fail in another aspect (e.g., grammatical ability). An analytic criterion allows the assignment of different scores to different subskills, thus the irregular development of the subskills in individuals can be graded accordingly (Hughes, 1989, Cohen, 1994).

Secondly, scorers are compelled to consider aspects of performance that they might otherwise ignore. Raters are required to assign a separate score for each aspect of a task that is stated in an analytic scale. If not stated separately, raters may consider different aspects of performance from each other or may overlook one or two aspects of performance, which may produce unfair results. Raters may be influenced by only one or two aspects of performance and assign their scores accordingly (Bailey, 1998, Cohen, 1994, Hughes, 1989, Madsen, 1983, Weir, 1995).

In addition, Weir (1995) directly states that "Analytic scoring can help provide a profile a student's weaknesses and strengths which may be helpful diagnostically, and also make a formative contribution in course design" (p. 45). Since students are placed on separate scales, each assessing a different aspect of performance such as grammar, vocabulary, and appropriateness, it is possible to explain why a particular score was assigned to each learner. The meaning of the score can be interpreted and student's weaknesses and strengths can be explain to other raters, students, teachers, and also parents (Bailey, 1998, Cohen, 1994, Heaton, 1990, Hughes, 1989, Madsen, 1983, Weir, 1995).

Another advantage is that the scorer has to give a number scores rather than a single score to a student and this will tend to make the scoring more reliable. Assigning a single score to the performance on the basis of an overall impression of it, as in holistic scoring, makes the outcome less reliable than with ratings including a series of scores. The fact of having certain number of bands and descriptors in each band at assessing the student's performance allows raters to assign more consistent scores to students and this should lead to greater reliability (Cohen, 1994, Hughes, 1989, Weir, 1995).

An analytic marking scheme is a more useful tool for the training of raters and the standardization of their ratings than is a holistic one. Training of raters is easier when there is an explicit set of analytic scales because an analytic scale offers raters the aspects of the performance that need to be considered with descriptors. Also, it is easier to explain why a particular score was assigned to a learner in analytic scoring whereas in holistic scoring its is not, since students are placed a single level on a scale (Bailey, 1998, Cohen, 1994, Heaton, 1990, Hughes, 1989, Madsen, 1983, Weir, 1995). In holistic scoring, Madsen (1983) claims that many teachers, especially those who are untrained in analyzing speech may find it difficult to evaluate many things simultaneously and to assign a single score on the basis of an overall impression of student's performance. An analytic scale guides raters to assign scores to certain components of the performance that is evaluated.

Disadvantages

There are also some problems associated with analytic scales. The main disadvantage of analytic scoring is the time that it takes because raters are required to consider the all aspects of performance and levels that are stated separately in the scale. Even with practice, analytic scoring takes longer than with the holistic method (Cohen, 1994, Hughes, 1989, Weir, 1995).

Hughes (1989) notes another disadvantage of analytic rating scales as " … the concentration on the different aspects may divert attention from the overall effect of the speech. In as much as the whole is often greater than the sum of its parts, a composite score may be very reliable but not valid" (p. 94). Raters may concentrate on the components of speech rather than overall communication.

Another disadvantage is that the scale may not be informative for learners, especially if the scale has neglected some aspect of performance. Since raters consider only the categories stated in an analytic scale, it is possible not to include all aspects of performance. For example, learners may wish to receive feedback on their ideas and organization, but actually find their grammar and vocabulary receive more attention by the teacher and/or rater (Cohen, 1994).

Cohen (1994) cites Hamp-Lyons' view that analytic scales may produce bias in favor of performances from which it is easiest to make judgments in terms of the scale. "This is why comments about grammar abound on essays - grammatical errors are some of the most external and easily accessible features of an essay" (Cohen, 1994, p. 318). Therefore, if analytic scoring is going to be used, aspects of performance need to be selected carefully and raters need to be trained to try to pay equal attention to all of them.

Constructing a Rating Scale

Rating scales for assessing productive skills have an essential place in achieving a high degree of reliability in a test. In order to measure the quality of spoken performance, first criteria of assessment need to be established.

During the stage of designing criteria for assessing the product of performance, decisions have to be made about how the performance will be judged, in other words, what to include in a test of spoken language. As tasks cannot be considered separately from the criteria that will be applied to the performances, the relationship between a task and criteria is an important issue in constructing rating scales. While constructing a rating scale, the theoretical definition of the construct to be measured and the test task specifications need to be considered. The way the construct for a particular test situation is defined determines which areas of language ability need to be scored. The way the test tasks are specified determines the type of performance that will be required of the learner. Of course, with performance assessments, there are many different possible ways for a test taker to respond. The rating scale must be broad enough to allow for all this possible performances and at the same time, specific enough so raters can judge each performance (Bachman & Palmer, 1996, McNamara, 1996, Weir, 1995).

After the areas of language ability to be assessed are defined, the scale definitions are specified. The scale definition includes two parts: the specific features of the language sample to be rated with the scale and the definition of scale levels in terms of the degree of mastery of these features (Bachman & Palmer, 1996).

Underhill (1987) indicates that how detailed the descriptor for each band should be is a problem in constructing a rating scale and states the following.

> The more information you give, the easier it will be for an assessor to find something that seems to match the learner sitting in front of her. At the same time, the more detail at each level, the more likely it is that some of it will be contradictory, or that statements in different categories will seem to place a learner at different levels (p. 99).

The question of how much detail needs to be given in scale definitions depend on the characteristics of the raters. Bachman & Palmer (1996) illustrates the situation in the following example.

> For example, if trained English composition teachers are rating punctuation, a construct definition that includes a list of all the punctuation marks in English may be unnecessary (p. 213).

Underhill (1987) also suggests keeping scale as simple as possible and not using more levels than needed. He notes that "The fewer levels you have, the easier it is to assess, and the higher the reliability will be" (p. 100).

According to Weir (1990), an assessment criterion can be developed and applied to samples of students' speech. The problem in developing criteria, especially for productive skills (speaking and writing), is that it is difficult to write explicit behavioral descriptions of levels within each of the criteria. Brindley (1998) points out that the writers of rating scales need to be very clear about the purpose which scales are meant to serve.

Underhill (1987) highlights the difficulty of designing rating scales and states that "The only solution is to adapt and improve the scales by trial and error, keeping only parts that are genuinely useful… Do not try to find the perfect scale" (p. 99). In order to find a scale that works well, it needs to be used and revised.

Training

After an appropriate assessment criterion is established, how best to apply the criteria to the samples of task performance needs to be considered. Although a standard criterion is used to assess oral ability, the scoring will be reliable only if scorers are trained to use the criterion (Weir, 1995).

As performance assessment typically involves judgment, the selection and training of raters is important (McNamara, 1996). Even if the examiners are provided with an ideal marking scheme there might always be some who do not mark in exactly the way required (Weir, 1990). Raters may have different expectations from learners or may differ in strictness in terms of assigning scores to learners.

Teacher training may influence teachers' assessment. Chalhoub-Deville (1996) cites research showing that in second language testing, trained teachers and non-teaching native speakers differ in their assessment of learners' second language oral ability. Consequently, assessment of learners' second language ability obtained from different groups may differ.

To reduce the variability of judges' behavior, raters should attend a training program in which they are introduced to the assessment criteria before assessing the learners. The training of examiners is seen as a crucial component of any testing program (Alderson, Clapham, & Wall, 1995, Bachman & Palmer, 1996, Cohen, 1994, Douglas, 2000, Hughes, 1989, McNamara, 1996, Underhill l987, Weir, 1990, 1995).

The purpose of standardization procedures is to bring examiners into line with each other and identify any factors which might lead to unreliability in marking and try and resolve these at the meeting so that candidates' marks are affected as little as possible by the particular examiner who assesses them (Weir, 1990).

During the training, the examiners need to become familiar with the marking system that they are expected to use and they must learn how to apply it consistently (Alderson, Clapham, & Wall, 1995). The raters are introduced to the assessment criteria and asked to rate a series of carefully selected sample performances. Sample performances illustrating a range of abilities and characteristic issues arising in the assessment are chosen. Ratings are carried out independently and after each performance is rated by all participants, raters are shown the extent to which they are in line with other raters. This leads to discussion and clarification of the criteria. The rating session is usually followed by additional ratings. This process is repeated for all of the selected performances. The procedure is used to determine whether the raters can participate satisfactorily in the rating process. After the standardization procedure examiners are allowed to assess candidates (McNamara, 1996, Weir, 1990).

"Until we can agree on precisely how speech is to be judged and have determined that the judgment will have stability, we cannot put much confidence in oral ratings" (Harris, 1969, p. 83). In oral testing, there is a need for explicit rating scales, and training and standardization of markers in order to boost test reliability (Weir, 1990).

Conclusion

This chapter has reviewed the literature related to this study. The next chapter will focus on the methodology, which covers the participants, instruments, procedures and data analysis used in the study.

CHAPTER 3: RESEARCH METHODOLOGY

Introduction

The objective of this research study is to investigate the inter-rater reliability of two alternative oral assessment criteria designed for Anadolu University, School of Foreign Languages. Teachers' perspectives on the use of the two alternative speaking assessment criteria will also be looked at. In order to be able to investigate the inter-rater reliability of two the different scoring systems, two sets of data were collected: raters' scores using both of the alternative oral assessment criteria and raters' opinions of the rating scales.

In this chapter, participants involved in the study, instruments used to collect data, data collection procedures and data analysis procedures are discussed in detail.

Participants

The participants involved in this research study are five English instructors currently employed at Anadolu University School of Foreign Languages. The participants were selected for the study on the basis of willingness to participate. The researcher explained the process of this research study to the instructors at Anadolu University and asked whether they would participate voluntarily in the study. Five of the instructors volunteered to participate in the study and signed the consent form (see Appendix B). One of the participants, rater 2 was excluded from the second workshop. She attended the training and norming sessions in the second workshop but could not grade the 36 students' oral performance because of a schedule conflict.

All of the participants are female and non-native speakers of English. The participants' ages ranged from 26 to 35. Their years of experience in teaching English ranged from three to eleven years. Among the five participants, four instructors were teaching speaking during the 2001-2002 fall and spring semesters. The other one had given speaking courses in the past. Their years of experience in assessing speaking ability ranged from three to eleven years.

Instruments

In order to look at the inter-rater reliability of the two alternative speaking assessment criteria, the following instruments are used: the two alternative rating scales, video recordings of 56 elementary level students, audiotape recordings of training and norming sessions, and group interviews, and the scores assigned by each participant to each student, using both of the alternative criteria.

The Two Alternative Rating Scales

The researcher developed two different rating scales to be used at Anadolu University School of Foreign Languages. Since the video recordings of speaking interviews were from elementary learners, the scales were developed to be used at the elementary level.

After reviewing the way speaking is assessed, the researcher developed an alternative criterion which is based on models from University of Cambridge Local Examinations Syndicate, Hughes, A. (1989) and Harris, D. P. (1969). The criterion is designed as an analytic criterion for three main reasons: First, an analytic criterion allows the assignment of different scores to different subskills, thus the irregular development of the subkills in individuals can be graded accordingly. Secondly, scorers are required to consider aspects of performance that they might otherwise ignore. Thirdly, the scorer has to give a number of scores for each category and this will tend to make the scoring more reliable (Hughes, 1989).

After deciding to have an analytic scale, the categories of the scale were determined. At this point, literature was taken into consideration. The construct of speaking ability and different kinds of rating scales were analyzed.

The goals and objectives of elementary level speaking classes at Anadolu University and the two criteria that are used for oral presentations and class participation scores were considered as the main sources for the new scales. The objectives of the speaking course at Anadolu University school of Foreign Languages are stated on the speaking course grading criteria document as follows:

> Students should be able to:
> - use structures and functions taught in speaking classess effectively
> - use vocabulary, idioms, expressions etc. taught in speaking classes
> - communicate and comprehend what is said and produce meaningful (formally, structurally and lexically appropriate) utterances.

Test tasks were also taken into account. Learners are required to perform two different tasks: a picture description and an information gap activity in which one learner is required to describe the locations of objects while the other learner listens to his/her partner and locates the object in the correct place on the picture. These tasks are chosen according to what is specifically taught in speaking classes. In conclusion, grammar, vocabulary, pronunciation, fluency and task achievement were chosen as the five categories in the scales.

The second step was to determine the number of bands in each category and to write descriptors for each category. Instructors employed at Anadolu University have been using a five-band speaking assessment scale to assess learners who are attending Open Education Faculty for three years. The fact that they are familiar with a five-band scale was also kept in mind. The sources mentioned above were again taken into consideration to decide the number of bands and descriptors. For each category five-band scale was chosen and the descriptors were written (see Appendix C).

Then the alternative criterion was e-mailed to seven teachers who are experts in English Language Teaching field to get feedback. According to the feedback the researcher received, the criterion was revised and the second alternative criterion was designed (see Appendix D).

The main difference between the first and second criteria is that the second alternative scale has four bands instead of five bands in each category. The reason for decreasing the number is to decrease the workload on raters. Since the criterion is analytic and has five categories with five bands and descriptors, decreasing the number of the bands is intended to help the raters to choose the appropriate band in each category for the student. As mentioned in the previous chapter, fewer bands will produce higher reliability (Underhill, 1987). Underhill (1987) also suggests not using more levels than are needed. The most problematic bands are the middle bands. It is always more difficult to decide to choose between the second or third band rather than the first band or the fifth band. The situation is same for choosing the third or the fourth band in a five-band scale. Having four bands for each category and the last band as " did not speak or spoke very little" instead of five solves the problem since there is only one middle band.

The second modification was to the descriptors of the scale. Some of the qualifiers in the five-band scale were not written in the four-band scale. The main point in each description was retained while the qualifiers were omitted. The sentences below are examples taken from the category of "vocabulary" in five-band and four-band scales to show the comparison in terms of the descriptors.

Five-band scale: VOCABULARY

5. Accurate and appropriate use of vocabulary with few noticeable wrong words, which do not affect communication

4. Occasional use of wrong words, which do not, however affect communication

3. Frequent use of wrong words, which occasionally may affect communication

2. Use of wrong words and limited vocabulary, which affect communication

1. Use of wrong words and vocabulary limitations (even in basic structures) result in disrupted communication

Four-band scale: VOCABULARY

3.Accurate and appropriate use of vocabulary with few noticeable wrong words

2.Use of wrong words occasionally may affect communication

1. Use of wrong words results in disrupted communication

0. Did not speak or spoke very little

The meanings of each descriptor were explained and discussed in detail during training and norming sessions. Each scale is intended to serve as a guide to help raters in assessing performance.

To sum up, both of the criteria are analytic and have the same five categories, which are grammar, vocabulary, intelligibility, fluency, and task achievement, with the same percentages assigned to each category. One difference between the two alternative assessment criteria is that one of them has five bands in each category while the other one has only four bands. In addition, the descriptors in the five-band criterion are more detailed.

Video recordings of elementary level students

Another instrument used in this study is videotape recordings of 56 elementary level students in the 1st speaking exam administered in 2001-2002 fall term.

After receiving permission from Preparatory School of Anadolu University administration to collect data, the researcher talked to teachers who were teaching speaking during 2001-2002 fall term and explained the study, including its aims, procedure, and future implications to our own institution. One of the speaking teachers who is teaching elementary levels agreed to help in collecting samples to be used in the study and she explained the study to the students in her classroom. Then the researcher sent a consent form to those elementary level students in the Preparatory School of Anadolu University and asked for their permission to videotape their first Speaking interview exams. In the consent form the purpose of the research study is explained (see Appendix E). The researcher met the students in one of their speaking courses and answered their questions related with the study. Fifty-six students signed the consent form.

Audio recordings of training and norming sessions

The training and norming sessions for teacher participants for both of the alternative criteria were tape recorded to be used in the data analysis. With the help of the analysis of the audio recordings, the problems that the participants had during these sessions were discovered and further implications for training and norming sessions could be suggested.

Audio recordings of the group interviews

Finally the participants were interviewed as a group after each workshop when the assessment of 36 students that are used in the study was completed. The interviews were used in order to obtain data for investigating participants' perceptions and attitudes toward speaking assessment in general, the scale they used in the workshop, and the training and norming sessions they received.

The interviews were held in Turkish. The audio recordings of the group interviews were transcribed and necessary portions were translated into English. The interviews were semi-structured. The interview questions used in the thesis investigating the reliability of the holistic grading system for the evaluation of the essays at the preparatory school of Eastern Mediterranean University in North Cyprus were taken as a model (Onurkan, 1999). In the first workshop, the researcher asked nine questions (see Appendix F), six of which were repeated in the second workshop as well (see Appendix G). The three questions unique to the first workshop covered problems in assessing oral performances and decision-making procedures for using scales in general. The six repeated questions asked about the training session the raters had received and the descriptors in the scale. The participants were asked follow-up questions to clarify or explain their ideas.

The interview questions focused on the problems that the participants faced while they were assessing learners' oral performances using the four-band and five-band scales. The aim of focusing on the problems is that the scale may be revised and some recommendations to solve the problems can be made.

Scores of each participant assigned to each student using both of the alternative criteria

In order to look at the inter-rater reliability of the two alternative oral assessment criteria, data were collected by means of having each instructor grade 36 students' oral performance. Statistical analysis was used to examine inter-rater reliability in the two alternative grading systems.

Procedures

Before proceeding with the research, I wrote a letter explaining the purpose of the research study and asked for permission for collecting data from in the Preparatory School of Anadolu University administration.

After receiving permission to collect data, the researcher sent a consent form to elementary level students in the Preparatory School of Anadolu University and asked for their permission to videotape their first Speaking interview exams. In the consent form the purpose of the research study was explained. The researcher met the students in one of their speaking courses and answered their questions related with the study. Fifty-six students signed the consent form.

The researcher found five instructors at Anadolu University who volunteered to participate in this research study. The participants were not told the focus of the study in order not to be affected. The researcher only explained the process of this research study to the participants.

Then the researcher designed an alternative speaking assessment criterion which is designed according to literature, the goals and objectives of elementary level speaking courses at Anadolu University, criteria for class participation and oral presentation used in elementary level speaking classes and test tasks that learners are required to perform during the interviews. Then the alternative criterion that the researcher developed was e-mailed to seven teachers who are experts in English Language Teaching field to get feedback about the criterion. According to the feedback the researcher received, the criterion was revised and the second alternative criterion was designed.

Next, the 56 elementary students were videotaped during their speaking exams. The tapes were used for the training, norming and study sessions. During the video recording of the speaking interviews, the conditions that may disturb the learners were minimized. For example, the camera was set up before the speaking exams started and there was not a camera operator in the class to record the students during the interview.

Then the sample student video recordings to be used during the training and norming sessions were chosen. While choosing sample interviews for the training, students illustrating different bands in the rating scales were chosen. For the norming session, problematic cases in which oral performance was not clear were included along with cases in which the performance could be scored in different bands in the criterion. Therefore the participants had a chance to discuss those students' scores and reach a consensus on how to deal with them.

Finally, the dates for the first and second workshop were decided with the participants. The first workshop was held on 6-7April, 2002 and the second workshop was held on 18-19 May, 2002.

The first step in workshop one was to have a training session on the use of the four-band speaking assessment criterion. The training session was held at Anadolu University School of Foreign Languages with the five participants. The participants were given the four-band speaking scale. Then eight sample student interviews (four pairs) illustrating different bands were introduced. The second step in workshop one was the norming session where the participants graded fourteen other sample interviews (seven pairs), which were chosen previously, using the four-band assessment scale. While the participants were assessing, the researcher observed them in case they needed help. The samples include performances illustrating different bands. In addition, performances that may be problematic to assess were also included so that the participants could have a chance to discuss them and reach a consensus. When they finished grading, the scores they gave for each student along with their reasons for giving them were discussed. The main reason for discussing the scores was to reach a consensus among the raters in using the scale. The training and norming sessions were tape recorded.

The third step in workshop one was that the participants assessed 36 video recordings of students the following day. The grading session was also held at Anadolu University School of Foreign Languages. The participants watched the recordings of 36 students on video and assessed them using the four-band speaking assessment scale.

After marking, the researcher interviewed the participants as a group. The participants were asked to express their opinions about speaking assessment in general, the training and norming sessions they received before marking in workshop one and also about the marking process.

Six weeks later, in workshop two, the five-band scale was used and the procedure was the same as in workshop 1. The main difference was that one of the raters, rater 2, was excluded from the second workshop. She attended the training and norming sessions but could not grade the 36 students' oral performance because of schedule conflict. In addition, all the video recordings of students were reordered so that the participants would not remember the students easily.

Data Analysis

The data analysis was performed in two steps. First, the scores given to 36 elementary students using the first alternative speaking assessment criterion by the five participants and the scores given to the same 36 elementary students using the second alternative speaking assessment criterion by the four participants were used and inter-class correlation for both of the alternative criteria was calculated.

Second, the interviews were analyzed by focusing on the problems that the raters faced while they were assessing learners' oral performances. The different interpretations of the two scales were uncovered. The common problems mentioned by the raters about the scales and assessing learners' oral performances were reported. The data analysis procedures and results will be explained in more detail in the following chapter.

CHAPTER 4: DATA ANALYSIS

Introduction

This study investigated the inter-rater reliability of two alternative oral language assessment criteria developed for Anadolu University School of Foreign Languages and also the participants' perspectives on the use of these two assessment criteria.

The chapter consists of two sections. In the first section, data analysis procedures are stated briefly. The second section covers the results of this research study. The second section has two sub-sections in which the qualitative and quantitative results are presented for each rating scale. The first section covers the results of the five-band speaking assessment scale, and the second section covers the four-band speaking assessment scale.

Presentation and Analysis of the Data

The study employed two sets of data: speaking scores using both of the assessment criteria and raters' opinions about the use of these two assessment criteria. The data collected from interview scores were analyzed quantitatively and are presented in tables. The data collected from raters' opinions were collected through interviews and the data gathered form the interviews were analyzed qualitatively.

In order to investigate the inter-rater reliability of the two alternative speaking assessment criterion two sets of data were collected: speaking scores given to 36 elementary students using both of the alternative speaking assessment scales and the participants' opinions about the use of these two scales.

The results of this research study are presented in four sub-sections. The first sub-section covers the quantitative results for the four-band speaking assessment scale. The second section looks at the qualitative results for the four-band scale. The third section gives the quantitative results for the five-band speaking assessment scale. Finally, the fourth section discusses the qualitative results for the five-band speaking assessment scale.

Inter-rater Reliability of Four-band Speaking Assessment Scale

In workshop one where the four-band speaking assessment scale used, the scores given to 36 elementary level students by five participants were collected. These scores are displayed in Table 1.

Table 1
The speaking scores given by the 5 raters using the four-band speaking assessment scale

Student Number	Rater 1	Rater 2	Rater 3	Rater4	Rater 5
1	97	87	90	77	74
2	100	97	100	100	90
3	94	100	94	94	94
4	81	84	90	84	71
5	97	100	100	100	74
6	49	55	58	59	65
7	45	48	39	42	52
8	67	80	68	61	71
9	74	80	100	90	90
10	68	68	58	68	71
11	42	39	39	36	39
12	39	36	42	36	39
13	49	52	39	52	68
14	71	71	65	58	81
15	71	58	71	64	71
16	55	58	71	64	71
17	58	56	53	71	71
18	59	65	65	65	71
19	100	100	100	100	100
20	100	100	100	100	100
21	87	90	74	81	90
22	100	90	81	74	80
23	90	90	71	71	84
24	90	83	100	71	84
25	80	87	94	74	90
26	68	68	78	80	71
27	77	80	68	100	100
28	68	73	75	81	59
29	42	45	45	42	71
30	58	58	77	52	68
31	62	62	68	52	65
32	71	68	78	71	71
33	59	59	75	88	88
34	49	59	59	55	59
35	36	36	65	36	49
36	39	36	49	36	49

In order to find out the inter-rater reliability of the four-band speaking assessment scale, the degrees of freedom (DF), sum of square (SM), mean square (MS) and Fischer's value (F) of the scores given by the five participants were calculated. The results are displayed in Table 2.

Table 2

Analysis of Variance for the four-band speaking assessment scale

Source	DF	SS	MS	F
Group 1	35	55502.2	1585.8	24.13**
Error	144	9461.6	65.7	
Total	179	64963.8		

Note. DF : Degrees of freedom, SS : Sum of square, MS : Mean square, F : Fischer's value
**: $p < .01$

The p-value obtained from this test is less than 0.01. As the null hypothesis is rejected when $p < .01$, the null hypothesis is rejected at the level of 0.00. The null hypothesis is as follows:

Null Hypothesis: There is no significant difference among the scores given by five participants.

According to the p-value, which is less than 0.01, it is concluded that there are significant differences among the scores given by the four participants. As seen in Table 1, the scores are not equal.

Then inter-class correlation (Winer, Brown, & Michels, 1991) was calculated using the results of the statistical analysis displayed in Table 2. Winer, Brown, & Michels (1991) define intra-class correlation as a measure of " … reliability within the context of the variance-component model of the analysis of variance" (p. 286).

According to the statistical analysis, it is concluded that the inter-rater reliability of the four-band speaking assessment scale is 0.82. As Brown (1996) puts it, the reliability coefficient can be interpreted as the percent of reliable variance in the scores on a test. Therefore, we can conclude that the scores are 82% consistent, or reliable, with 18% measurement error (100%-82%=18%), or random variance.

The recommended reliability coefficient for oral production tests is in the .70 to .79 ranges, which is considered adequate for oral tests (Brown, 1996, Hughes, 1989, Lado, 1961). Consequently, we can conclude that the reliability of the four-band scale is over .70 to .79 ranges and can be considered more that adequately reliable with the .82 reliability coefficient.

Raters' opinions about the four-band speaking assessment scale

In order to get information about the four-band speaking assessment scale, the participants were interviewed about the scale after the grading session finished. The interview was conducted as a group interview.

The interview results indicate that the raters have some common problems concerning the descriptors used in the criterion. The problems that the raters face in the scoring procedure can be grouped under five headings: 1) one term used in the descriptors is not clear, 2) students' performance may not fit into the bands, 3) the number of bands in each category is not enough, and the highest band in vocabulary needs to be more detailed 4) the lowest band is unnecessary, 5) there is a big difference among the bands in terms of the value assigned to each band.

The raters expressed their ideas on the problems they faced while grading the students' oral performances in the following ways.

1) One term used in the descriptors is not clear

Three participants out of five stated that one term used in the descriptors was not clear for them. The below sentences are examples taken from the interviews.

> Rater1: ... well...actually we discussed the term *few* during the norming session. We couldn't agree on what it means.
>
> Rater 3: ... yeah. What we understand can change. We can interpret the band differently.
>
> Rater 4: …yes. What "few" means to me may not mean the same thing to what it means to you.

The above sentences taken from the interviews held with the five raters showed that one term used in the descriptors was not clear for them. In addition, the interview results show that the raters believed they might interpret the band that includes the term *few* differently. This may result in assigning inconsistent scores to learners.

As the language of rating scales includes qualifiers such as sometimes, most, often, occasionally, the selection and preparation of raters is important (McNamara, 1996, Underhill, 1987, Weir, 1990). Weir (1990) states that the purpose of standardization procedures is to bring examiners into line and identify any factors which might lead to unreliability in marking and try and resolve these at the meeting so that candidates' marks are affected as little as possible by the particular examiner who assesses them. The procedure is used to determine whether the rater can participate satisfactorily in the rating process. After the standardization procedure examiners are allowed to assess candidates (Alderson, Clapham, & Wall, 1995, Weir, 1990, McNamara, 1996).

2) Students' performance may not fit into the bands.

Four raters out of five stated that they had difficulty in choosing the appropriate band for the students' performance. The sentences below are examples taken from the interview.

Rater 1: ...For example, the performance is described in the bands. The description may not fit the students' real performance.

Rater 5: ….I wholly agree with you, rater 1.

Rater 3: ...yeah the descriptors and the performance do not match and then what do you do? You try to choose the closest band in meaning. Actually, at that point, the band does not describe the student's performance but you choose that band.

Rater 4: …yeah, I agree. And at that point, how can you be objective?

The above raters stated that they sometimes had problems in matching a student's performance with a band. They claimed that they sometimes had problems while scoring the performances because some of the performances did not fit into the descriptors.

As one of the raters mentioned, if the descriptors and the learner's performance do not match, you try to choose the closest band in meaning. This is exactly how banded scales work. According to Underhill (1987), the rating scale offers the assessor a series of prepared descriptions and then he or she picks the one which best fits each learner. The raters were provided with this information during the training sessions.

In order to be consistent in choosing the appropriate band for each learner by all the raters, that raters need to be trained and standardized. During the training and standardization, the raters become familiar with the marking system that they are expected to use and they must learn how to apply them consistently (Alderson, Clapham, & Wall, 1995).

3) The number of bands in each category is not enough and the highest band in vocabulary needs to be more detailed.

Five raters stated that they had difficulties in grading the students' oral performances because the number of bands in each category was not enough. The sentences below are examples taken from the interview.

Rater 2: … I think there can be some more bands

Rater1: ... I agree with you, rater 2.

Rater 3: ...well, that's logical to decrease the number of the bands to four to assign the scores fast but there is a difference between assigning the scores fast and assigning the grade that the learner deserves.

Rater 1: … well, I don't support the idea of having this number of bands because the more you have bands the more performance you describe.

Rater 4: I agree with you, rater 1. We need more bands that describe students' performance.

Rater 5: ...and for example, I had to assign the same scores to two students but the students' performances are different from each other and two different performances got the same score.

The above sentences taken from the interviews held with the five raters showed that they felt the number of bands in each category was not enough.

Four raters out of five stated that they sometimes had problems in the highest band in vocabulary since it was not detailed enough.

Rater 1: … we got a problem in vocabulary yesterday in the norming session. In vocabulary, between the first and the second bands. We could not differentiate them.

Rater 4: I think we need to add something to the first band.

Rater 2: The first band should say "use of noticeable wrong errors which do not affect communication". Because some students used wrong words and it affected the meaning. And some students used wrong words and it did not affect the meaning that much.

Rater 3: We want that band more detailed because there are some wrong words which affect communication and which do not affect communication.

As it is seen in the above sentences the four raters stated that they wanted the highest band in vocabulary be more detailed in order not to have difficulties while assessing the learners' oral performance. This would apply for the grammar as well since the descriptors in grammar and vocabulary are similar (see Appendix D).

Although the participants of this study stated that the number of the bands was not enough, they chose appropriate bands and assigned consistent scores to learners (see Table 6). Underhill (1987) supports the idea of keeping the scale as simple as possible and claims that "The fewer levels you have, … the higher the reliability will be" (p. 100). He also adds that one of the problems while designing scales is how detailed the profile for each learner should be and states that the more information you give in the scale, the more likely it is that some of the categories, levels, or statements will be contradictory.

4) The lowest band is unnecessary.

The five raters stated that they had difficulties in using the lowest band because every student produces something and did not fit into the lowest band. The sentences below are examples taken from the interviews.

Rater5: ….let me mention this point also… The band, which says, "didn't speak or spoke very little" is unnecessary because we force students to speak

Rater 2: [we give many prompts to students to produce something…

Rater 3: [yes..

Rater 4: [There are very very few students who spoke very little. For example, if the student fit into that band we as teachers give prompts to the student and he fits into another band.

Rater 1: … yeah… Then the student gets 2 rather than 1. We could never give the lowest band. We just assigned it to one student, as far as I remember.

The above five raters stated that they had problems using the lowest band because teachers gave many prompts to students during the interview.

In this situation, two different problems are mentioned. The first one is about giving prompts to students during the interview. Examiners may differ in terms of giving prompts to students. The problem is directly related with how the interview needs to be conducted. This suggests that the examiners also need training in how to conduct the interview consistently (Cohen 1980, Harris, 1969, Hughes, 1989, Weir, 1990).

The second problem is about the use of the lowest band. The raters stated that there were very few students who fit the lowest band and therefore that band was not used very often. Brown (1996) mentions that because the achievement tests are designed with very specific reference to a particular course and directly based on course objectives, "a good achievement test can tell teachers a great deal about their students' achievements and about the adequacy of the course" (p. 14). Since the scale is designed and used for an achievement test, which is administered at the end of the first term, it is quite normal not to have students who fit the lowest band. If the test is administered at the beginning of the term, there should be some students who would fit the lowest band.

5) There is a big difference among the bands in terms of the value assigned to each band.

Three raters out of five stated that there was a big difference among the bands in terms of the value assigned to each band.

> Rater 4: … also between the points assigned to each band… well if you compare our scores most probably you will see this: for example, one of us gave 30 and another rater gave 20. This makes 10 point difference, a big difference.
>
> Rater 5: … if the same thing happens in grammar as well… then the difference will be 20 points.
>
> Rater 2: … also for example, yesterday, one of the students got 55 and the other got 97. But there was a very slight difference between them in terms of their performance. This is because of the bands, because of the ten-point difference.

The above sentences taken from the interviews showed that there could be big differences among the raters because of the value assigned to each band. They also stated that there could be big differences among the students since there is a 10-point difference between the bands in grammar, vocabulary categories.

The example given, of students' scores of 55 and 97, is from the norming sessions in which the raters are supposed to discuss the scores assigned to each learner and reach a consensus among themselves. This is not a problem in the norming session. After the training and norming sessions, raters should not have this kind of problem. The sample interviews for training and norming sessions are chosen beforehand. While choosing sample interviews for the training, students illustrating different bands were chosen. For the norming session, problematic cases in which oral performance was not clear were included along with cases in which the performance can be scored with high, mid and low bands in the criterion. Therefore the participants had a chance to discuss about those students' scores and reach a consensus on how to deal with them. This process is repeated for all of the selected performances in the norming session. The procedure is used to determine whether the rater can participate satisfactorily in the rating process. After the standardization procedure examiners are allowed to assess candidates (McNamara, 1996, Underhill, 1987, Weir, 1990). Raters need to attend the training and norming sessions not to have problems in using the scale consistently. The quantitative analysis of the data here reveals that raters produced more than adequately reliable scores.

Inter-rater Reliability of Five-band Speaking Assessment Scale

In workshop two where the five-band speaking assessment scale used, the scores given to 36 elementary level students by four participants were collected. These scores are displayed in Table 3.

Table 3
The speaking scores given by the 4 raters using the five-band speaking assessment scale

Student Number	Rater 1	Rater 3	Rater4	Rater 5
1	84	84	88	88
2	84	94	88	88
3	76	86	84	88
4	80	86	70	70
5	100	100	100	100
6	86	84	74	86
7	48	48	50	42
8	68	64	62	68
9	100	100	90	94
10	96	84	84	88
11	68	62	64	46
12	60	72	50	50
13	74	72	82	74
14	72	84	86	86
15	66	80	68	66
16	68	78	74	68
17	78	78	78	68
18	70	74	74	68
19	100	100	100	100
20	100	100	100	100
21	82	82	88	72
22	94	96	80	78
23	86	88	82	88
24	80	92	80	86
25	80	98	74	80
26	72	78	64	62
27	86	90	82	80
28	80	76	80	74
29	72	82	74	58
30	80	74	70	64
31	70	74	74	76
32	68	70	82	78
33	86	86	80	74
34	62	50	42	58
35	60	56	44	50
36	56	60	56	44

In order to find out the inter-rater reliability of the five-band speaking assessment scale, the degrees of freedom (DF), sum of square (SM), mean square (MS), and Fischer's value (F) of the scores given by the five participants were calculated. The results are displayed in Table 4.

Table 4

Analysis of Variance for the five-band speaking assessment scale

Source	DF	SS	MS	F
Group 2	35	29233.00	835.23	9.94**
Error	108	9073.00	84.01	
Total	143	38306.00		

Note. DF : Degrees of freedom, SS : Sum of square, MS : Mean square, F : Fischer's value
*: p < .05

The p-value obtained from this test is less than 0.01. As the null hypothesis is rejected when p < .01, the null hypothesis is rejected at the level of 0.00. The null hypothesis is as follows:

Null Hypothesis: There is no significant difference among the scores given by four participants.

According to the p-value, which is less than 0.01 it is concluded that there are significant differences among the scores given by the four participants. As it is seen in Table 3, the scores are not equal.

Then inter-class correlation (Winer, Brown, & Michels, 1991) was calculated using the results of the second statistical analysis displayed in Table 4. The finding of the inter-class correlation of the five-band speaking assessment criterion is 0.69.

According to the statistical analysis, it is concluded that the inter-rater reliability of the five-band speaking assessment scale is 0.69. As seen earlier, a reliability coefficient of .70 to .79 is adequate for oral production tests (Brown, 1996, Hughes, 1989, Lado, 1961). Consequently, we can say that the five-band scale is close to .70 reliability coefficient but cannot be considered adequately reliable with the .69 reliability coefficient.

Raters' opinions about the five-band speaking assessment scale

In order to get information about the five-band speaking assessment scale, the participants were interviewed about the scale after the grading session finished. The interview was conducted as a group interview.

The interview results indicate that the raters had a common problem concerning the descriptors used in the criterion. The problem that the raters faced in the scoring procedure can be grouped under this heading: two terms used in the descriptors are not clear.

The raters expressed their ideas on the problem they face while grading the students' oral performances in the following ways.

All of the raters stated that terms used in the descriptors were not clear for them. The below sentences are examples taken from the interviews.

Rater 5: … well we got problems with some terms: occasionally and sometimes, while assigning scores to the interviews

Rater1: [the descriptors which say sometimes and occasional caused some problems

Rater 4: [yeah... if we look at the second part of the descriptor which talks about errors, which says "which may affect" or "which do not affect", then it becomes more clear.

> Rater 3: We had difficulty with the terms *occasional and sometimes* but we discussed the terms during the norming session and tried to solve the problem.

The four raters claimed that they sometimes had problems with two terms used in the descriptors since they were not clear for them. The terms are *occasional* and *sometimes* that take place in some of the descriptors in the scale.

As noted in the discussion of the four-band rating scale, qualifiers are a common problem in performance assessment. The raters may interpret the scale differently as the terms are not clear for them and assign inconsistent scores to learners.

The solution for this problem is training and standardizing the raters. According to Alderson, Clapham, & Wall (1995) and McNamara (1996), in order to reduce the variability of judges' behavior, raters attend a training program in which they are introduced to the assessment criteria before assessing the learners. The training of examiners is seen as a crucial component of any testing program. In addition, Weir (1990) states that the purpose of standardization procedures is to bring examiners into line and identify any factors which might lead to unreliability in marking and try and resolve these at the meeting so that candidates' marks are affected as little as possible by the particular examiner who assesses them.

As it is seen in the above sentences the raters stated only one problem with the five-band scale. One reason may be the assessment scale since all the raters have been using a similar scale, which has five bands, for three years in the Open Education program to assess learners' oral performances.

In this chapter, the data collected from interview scores assigned to 36 elementary level students using the five-band and four-band scales and raters' opinions about these two scales were analyzed and interpreted. In the next chapter, the results will be further discussed.

CHAPTER 5: CONCLUSION

Overview of the study

This study determined the inter-rater reliability of the two alternative speaking assessment system developed for Anadolu University School of Foreign Languages and also the participants' perspectives on the use of these two assessment systems.

In order to achieve this purpose, two sets of data were collected: speaking scores using both of the assessment systems and raters' opinions. The participants were five English instructors currently employed at Anadolu University School of Foreign Languages. The participants attended the training and norming sessions for the five-band speaking assessment scale and graded 36 elementary level students' oral performance using the scale. One and a half months later, four of the participants attended the training and norming sessions for the four-band speaking assessment scale and graded the same 36 elementary level students' oral performance using the scale. The participants were asked to express their opinions about both of the scales following the marking process. The data were analyzed in two stages. First, inter-class correlation for both of the alternative speaking assessment criteria was calculated using the scores assigned to 36 elementary level students. Second, the interviews were analyzed by focusing on the theme being investigated.

General Results

This section discusses the findings and the conclusions that have been drawn through the process of data collection in order to answer the research questions. Each sub-section refers to one research question.

1. What is the inter-rater reliability of the four-band speaking assessment scale developed to be used at Anadolu University, School of Foreign Languages?

The result of the inter-class correlation (Winer, Brown, & Michels, 1991) of the four-band speaking assessment scale is 0.82. Therefore, we can conclude that the scores are 82% reliable.

Consequently, we can conclude that the reliability of the four-band scale is over the .70 to .79 ranges, which is considered adequate for oral tests (Brown, 1996, Hughes, 1989, Lado, 1961), and can be considered more than adequately reliable with the .82 reliability coefficient.

2. What are the participants' perspectives on the use of the four-band speaking assessment scale?

According to the interview results held with the five participants on the use of the four-band speaking assessment scale, the raters have some common problems concerning the use of the scale.

The three participants out of five stated that one term, *few,* used in the descriptors is not clear for them. As mentioned above, this is a common problem with the rating scales and the solution to clarify the terms for the raters is the same, which is training and standardizing the raters.

Also, the four raters out of five stated that they had difficulty in choosing the appropriate band for the students' performance. They stated that students' performance may not fit into the bands. The solution for this problem is again training and standardizing raters on the use of the scale so that they can choose the band, which best fits the learner's performance, consistently. (Alderson, Clapham, & Wall, 1995, Underhill, 1987)

The interview results also revealed that the number of the bands in the scale and the descriptor in the highest band in vocabulary are not enough and need to be more detailed. Since the descriptors in vocabulary and the grammar are similar, this concern applies for the grammar band as well (see Appendix D). This is a common problem in designing rating scales. Underhill (1987) suggests keeping the scale as simple as possible. The more descriptors or bands you have, the more difficult it is to assess and the lower reliability will be.

Another problem, which the raters faced, was related with the lowest band. The raters stated that they do not feel like they are using the lowest band. There are two reasons for the problem mentioned by the raters. First, the teachers give more prompts than they should give during the interviews, which is a problem of training raters on how to conduct the interviews. Second, since this is an achievement test, which is administered at the end of the course or program, test tasks directly focus on what is taught and studied in speaking classes, (Brown, 1996). Therefore, most of the students perform the tasks. In brief, raters also need training on how to conduct interviews, and it is quite normal not to have few students who fit the lowest band in achievement tests.

Finally, the interview results showed that there can be a big difference between the learners’ scores although there was a slight difference in terms of their performances. Three raters out of five stated this problem. In order to overcome the problem, raters need to a reach consensus to choose the band which best fits each learner’s performance among themselves with the help of the training and norming sessions.

3. What is the inter-rater reliability of the five-band speaking assessment scale developed to be used at Anadolu University, School of Foreign Languages?

The result of the inter-class correlation (Winer, Brown, & Michels, 1991) of the five-band speaking assessment scale is 0.69. Therefore, we can conclude that the scores are 69% reliable.

Consequently, we can say that the five-band scale is close to .70 reliability coefficient, which is considered adequate for oral tests (Brown, 1996, Hughes, 1989, Lado, 1961), but cannot be considered adequately reliable with the .69 reliability coefficient. With more training, the reliability coefficient of the five-band scale might be increased.

4. What are the participants' perspectives on the use of the five-band speaking assessment scale?

The interview results held with the four participants on the use of the five band speaking assessment scale indicated that the raters face one problem on the use of the scale. One reason why the participants mentioned only one problem may be the scale itself since they have been using a five-band scale for three years for the open Education oral interviews. This may lead the raters to have fewer issues to deal with the five-band scale.

The four raters claimed that they sometimes have problems with two terms used in the descriptors since they are not clear for them. The terms are *occasional* and *sometimes* that take place in some of the descriptors in the scale. As Underhill (1987) states that this is a common problem because qualifiers such as sometimes, most, often, occasionally, are a common feature of the language of rating scales. The solution for this problem is training and standardizing the raters on the use of the scale so that they can interpret the terms used in the scale consistently among themselves.

Discussion

The results of this research study revealed that the reliability of the four-band scale is over 0.70 to 0.79 ranges, which is considered adequate for oral tests (Brown, 1996, Hughes, 1989, Lado, 1961). Also, Hamp-Lyons (1990) state that as a result of many studies the score reliability has been raised around 0.80, which is commonly regarded as a satisfactory level for decision-making purposes, and has been stabilized. Therefore the four-band scale can be regarded as satisfactory for speaking achievement tests administered at Anadolu University School of Foreign Languages and considered highly reliable with the 0.82 reliability coefficient.

When compared to the five-band scale, the four-band scale proves to be more reliable. With the limited training and norming sessions that the five raters received, the statistical analysis showed that the four-band speaking assessment scale is highly reliable, in terms of inter-rater reliability. With more training and norming sessions, the reliability coefficient might be increased.

The five-band scale is close to .70 reliability coefficient, which is considered adequate for oral tests (Brown, 1996, Hughes, 1989, Lado, 1961), but cannot be considered adequately reliable with the .69 reliability coefficient. With more training and standardization, the inter-rater reliability coefficient might be increased. However, it is not practical to train and norm 82 teachers who take part in the scoring of oral assessment at Anadolu University School of Foreign Languages.

Almost 2000 students are assessed in the oral interviews and all of the teachers are participants in the oral interviews and scoring procedure. Since there are

82 teachers currently employed at the university, the scale that produces reliable scores with minimum amount of training and norming sessions is needed for the sake of practicality. Teachers have at least 18 hours of workload including two hours of substition in a week. Teachers also take part in test preparation and assessment for the skills they are teaching. In addition, the institution does not have the necessary resources needed for training and norming sessions. In each mid-term exam, different test tasks are chosen for a level according to the syllabus. Also, different test tasks in different levels are used in the interviews. Therefore, video recordings of student interviews from different levels to are needed to use in the sessions. Because of the reasons mentioned above and time constraints, it is difficult to conduct training and norming sessions for 82 teachers.

Consequently, we can say that the results of this research study suggests the use of four-band scale since the raters can assign more reliable, consistent scores to learners while evaluating their oral performances with minimal training and norming.

According to the results of this research study, and for the reasons mentioned above, the four-band speaking assessment scale is suggested for use in grading elementary level speaking interviews which are administered as achievement tests at Anadolu University School of Foreign Languages.

Recommendations

According to the results of the study, the four-band speaking assessment scale, which is developed for elementary level students at Anadolu University School of Foreign Languages, is recommended.

Raters mentioned problems about the way the interviews conducted. They stated that some teachers give more prompts than the others and this may affect learners' performances. McNamara (2000) states "There needs to be an agreement about the conditions (including the length of time) under which the person's performance or behavior is elicited, and/or is attended to by the rater" (p. 36). Therefore, the instructors who are going to take part in the scoring procedure need training in how to conduct the interview consistently. (Cohen, 1980, Harris, 1969, Hughes, 1989, Weir 1990).

As stated in Chapter 2, training and standardizing the raters play an important role to overcome the problems stated by the participants during the interviews. To reduce the variability of judges' behavior, the training of examiners is seen as a crucial component of any testing program (McNamara, 1996, 2000, Alderson, Clapham, & Wall, 1995). The teachers who are going to take part in the assessment of learners' oral performances need to go through more interviews and meet certain standards before they take part in the actual scoring procedure.

After an analysis of the two alternative analytic speaking assessment criteria, some recommendations can be made for the improvement of the two alternative speaking assessment criteria. Since nearly all the raters stated the problem of qualifiers used in the bands in both of the scales, the terms need to be reconsidered and paid more attention to during training and norming sessions. Recommendation for the improvement of the two criteria includes revising and clarifying some of the terms used in the descriptors such as *occasionally, sometimes, and few* in order to make them clear for the raters.

It is also recommended that the weighting of the categories may need to be revised. The participants mentioned the problem of big differences in terms of the values assigned to each category; therefore, revising the weighting needs to be considered.

Limitations

The major limitation of this research study is conducting it only with a limited number of raters who are non-native, female instructors. Four raters for the five-band speaking assessment scale and five raters for the four-band speaking assessment scale participated in the study. The more raters participating in the study, the easier it would be to generalize the results. More raters, including native and male speakers may have brought further insights to the results investigated in this study.

Another limitation is about the group interviews held at the end of each workshop to find out raters' opinions about the two alternative speaking assessment criteria. Raters may be affected by each other or may not have wanted to disagree with other participants. If they had been interviewed individually, the results of the qualitative analysis might have been different. One aim of interviewing the participants as a group is that they need to be interviewed just after the marking process. It might have taken two or more days to interview them as they are busy in the weekdays. Second, since both of the workshops were held in two days and lasted nearly four or five hours in a day, it was practical to interview them as a group in terms of time constraints.

Implications for Further Research

Those interested in further research might collect more data from more raters, including native and male speakers of English in order to see whether those raters have similar problems with the two alternative criteria.

With the data collected in this study, another research question can be investigated if the scores given to students are analyzed quantitatively: In the use of which scale are learners assigned higher scores and therefore considered more successful? The results of the study may provide valuable information for the use of the scales.

The validity of the sub-skills in the four-band scale which is suggested to be used for elementary level students can also be looked at with the data collected in this research study. Whether the categories in the scale assess same or different concepts can be investigated and recommendations can be made for the categories.

Another research study, which looks at inter-rater reliability of raters within each category in the four-band assessment scale, can be conducted. The scores given for each category by a number of raters can be analyzed quantitatively. The results of the study may provide useful information for the use of the scale. The categories which raters may have inadequate inter-rater reliability can be investigated and some recommendations for the categories can be made to improve the scale.

In addition, the alternative scales are designed for elementary level at Anadolu University School of Foreign Languages. Other scales for other levels taking the four-band scale as a model need to be designed to have consistency among levels. Also, data from other levels need to be collected to look at the inter-rater reliability of the scales.

Also, more data can be collected to investigate the intra-rater reliability of the four-band scale. At Anadolu University School of Foreign Languages, two raters assess learners' oral performance in the interviews in order to increase the reliability of the assessment. Therefore, the results of a study investigating the intra-rater reliability of the four-band scale may bring useful insights for the effectiveness of the scale.

Finally, as the raters mentioned that they had some problems while using the suggested four-band scale, studies on teachers' attitudes toward the scale in the near future will be helpful to see whether there will be any changes on the use of the scale.

Conclusion

This research study investigated whether there are significant differences among raters in their use of two alternative speaking assessment system developed for Anadolu University School of Foreign Languages and also the participants' perspectives on the use of these two assessment systems.

The results were drawn from two types of data: speaking scores using both of the assessment systems and raters' opinions. According to the results of the study, the four-band speaking assessment scale is suggested to be used at Anadolu University School of Foreign Languages to assess elementary level students' oral performances.

REFERENCES

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Brindley, G. (1998). *Describing language development? Rating scales and SLA*. In Bachman, L. F. & Cohen, A. D. (Eds.). *Interfaces between second language acquisition and language testing research.* Cambridge: Cambridge University Press.

Brown, J. D. (1996). *Testing in language programs.* Upper Saddle River, NJ, USA: Prentice Hall Regents.

Chalhoub-Deville, M. (1996). Performance assessment and the components of the oral construct across different tests and rater groups. In M. Milanovic.& N. Saville (Eds.) *Studies in language testing 3.* Cambridge: Cambridge University Press.

Clark, J. L. D. (1972). *Foreign language testing: Theory & practice*. Philadelphia: The Center for Curriculum Development, Inc.

Cohen, A. D. (1994). *Assessing language ability in the classroom.* Boston: Heinle & Heinle Publishers.

Davies, A. (1990). *Principles of language testing*. Oxford: Basil Blackwell Ltd.

Douglas, D. (2000). *Assessing languages for specific purposes.* Cambridge: Cambridge University Press.

Edelman, C. (1987). Effective interviewing in the communicative classroom. *Forum 25*, 3, 12-16.

Engelskirchen, A., Cottrell, E., & Oller, J. W. (1981). A study of the reliability and validity of the Ilyin Oral Interview. In A. S. Palmer, P. J. M. Groot, & G. A. Trosper (Eds.) *The construct validation of tests of communicative competence.* Washington, D.C.: Teachers of English to Speakers of Other Languages.

Gronlund, N. E. (1998). Assessment of student achievement. London: Allyn and Bacon.

Hamp-Lyons, L. (1990). *Second language writing: Assessment issues*. In Kroll, B. (Ed.). *Second language writing*. Cambridge: Cambridge University Press.

Harris, D. P. (1969). *Testing English as a second language*. New York: McGraw-Hill Book Company.

Heaton, J. B. (1990). *Classroom testing*. New York: Longman Group Limited.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Jones, R. L. (1979). The FSI oral interview. In Spolsky B. (Ed.) *Advances in language testing series*. Arlington, VA, USA: The Center for Applied Linguistics.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests.* London: Longman

O'Malley, J. M. & Pierce, L. V. (1996). *Authentic assessment for English language learners.* no city: Addison-Wesley Publishing Company.

Onurkan, G. (1999). *The reliability of the holistic grading system for the evaluation of essays at the preparatory school of Eastern Mediterranean University in North Cyprus.* Unpublished master's thesis. Bilkent University Ankara, TURKEY.

Madsen, H. S. (1983). *Techniques in testing*. Oxford: Oxford University Press.

McNamara, T. (1996). *Measuring second language performance*. London: Longman.

McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press.

Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision-making behaviour of composition markers. In Milanovic, M.& Saville, N. (Eds.) *Studies in language testing 3.* Cambridge: Cambridge University Press.

Shohamy, E. (1981). Inter-rater and intra-rater reliability of the oral interview and construct validity with cloze procedure. In A. S. Palmer, P. J. M. Groot, & G. A. Trosper (Eds.) *The construct validation of tests of communicative competence.* Washington, D.C.: Teachers of English to Speakers of Other Languages.

Underhill, N. (1987). *Testing spoken language*. Cambridge: Cambridge University Press.

Weir, C. J. (1990). *Communicative language testing.* London: Prentice Hall International.

Weir, C. J. (1995). *Understanding & developing language tests*. New York: Phoenix ELT.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design.* Boston: McGraw-Hill Inc.

APPENDIX A

Speaking examinations: assessment criteria

	Task achievement	Fluency	Accuracy and appropriacy
100 95 90 85	Tasks dealt with fully and effectively	Able to sustain flow of language appropriate to the tasks, with natural speed and hesitation	Generally effective use of structure, vocabulary and paraphrase at this level
80 75 70	Tasks dealt with adequately and effectively	Minimal hesitation to search for language	Basic structures and vocabulary are used appropriately. Difficult structures may sometimes be inaccurate. Few use of inappropriate vocabulary.
65 60 55	Tasks dealt with mainly adequately and effectively with some inadequacies.	Noticeable hesitations but not such as to strain the listener or impede communication	Meaning is conveyed despite noticeable inaccuracies in basic structures, lack of basic vocabulary and ineffective paraphrase
50 45 40	Limited ability to deal with tasks.	Hesitation often demand unreasonable patience of the listener	Meaning occasionally obscured by structural inaccuracies and/or limited vocabulary and inability to paraphrase
35 30 25	Ineffective handling of tasks	Speech very disconnected and difficult t follow	Frequently incomprehensible because of limited vocabulary and numerous structural errors
20 15 10	Unable to deal with tasks	No connected speech	Virtually incomprehensible because of insufficient vocabulary and gross structural errors

5 Students attended the interview, but did not speak.

APPENDIX B

INFORMED CONSENT FORM FOR PARTICIPANTS

Dear Participant:

You are being asked to participate in an experimental study. One aim of the study is to test the inter-rater reliability of two alternative oral assessment criteria designed to be used at Anadolu University School of Foreign Languages. In addition, your perceptions about the two alternative assessment criteria will be found out. In order to obtain the required data, you are being asked to assess learners' oral performances and state your opinions during the group interview. The details will be explained to you in the workshops which will be conducted by the researcher.

Your participation in this study will bring invaluable contributions to the study, and hopefully, own program. Any information given to the researcher will be kept confidential. This study involves no risk to you.

I would like to thank you for your participation in advance. If you have any questions please do not hesitate to contact me at the e-mail address given below.

Very Truly Yours,

ECE SELVA KARSLI
MA TEFL Program
Bilkent University
Ankara

e-mail: eskarsli@anadolu.edu.tr

I have read and understood the information given above. I hereby agree to participate in the study.

Name:

Date:

Signature:

APPENDIX C

FIVE-BAND SPEAKING ASSESSMENT SCALE

GRAMMAR 30		
5.accurate and appropriate use of grammar with few noticeable errors which do not affect communication	**30**	
4.occasional use of grammar errors which do not, however, affect communication	**24**	
3. frequent use of grammar errors which occasionally may affect communication	**18**	
2. use of grammar errors which affect communication	**12**	
1. use of grammar errors (even in basic structures) result in disrupted communication	**6**	
VOCABULARY 30		
5. accurate and appropriate use of vocabulary with few noticeable wrong words which do not affect communication	**30**	
4. occasional use of wrong words which do not, however affect communication	**24**	
3. frequent use of wrong words which occasionally may affect communication	**18**	
2. use of wrong words and limited vocabulary which affect communication	**12**	
1. use of wrong words and vocabulary limitations (even in basic structures) result in disrupted communication	**6**	
INTELLIGIBILTY (20)		
	20	
5. easily understandable	**16**	
4. little difficulty in being understood	**12**	
3. occasional difficulty in being understood	**8**	
2. frequent difficulty in being understood	**4**	
1.difficult to understand		
FLUENCY 10		
	10	
5.natural flow of speech with minimal hesitation	**8**	
4.occasional hesitation, which do not interfere with communication	**6**	
3.frequent hesitations, which occasionally may affect communication	**4**	
2.usually hesitant that affect communication	**2**	
1.no connected speech result in disrupted communication		
TASK ACHIEVEMENT 10		
	10	
5.tasks completed fully	**8**	
4. tasks completed adequately	**6**	
3. tasks completed almost adequately	**4**	
2. tasks completed inadequately	**2**	
1. tasks not completed		
TOTAL		

APPENDIX D

FOUR-BAND SPEAKING ASSESSMENT SCALE

VOCABULARY (30) 3- accurate and appropriate use of vocabulary with few noticeable wrong words 2-use of wrong words occasionally may affect communication 1-use of wrong words results in disrupted communication 0-did not speak or spoke very little	**30** **20** **10** **1**	
GRAMMAR (30) 3-accurate and appropriate use of grammar with few noticeable errors 2-errors of grammar occasionally may affect communication 1-errors of grammar result in disrupted communication 0-did not speak or spoke very little	**30** **20** **10** **1**	
INTELLIGIBILTY (20) 3-easily understandable 2-occasional difficulty in being understood 1-difficult to understand 0-did not speak or spoke very little	**20** **14** **8** **1**	
FLUENCY (10) 3-natural flow of speech with minimal hesitation 2-hesitation that occasionally may affect communication 1-hesitations that result in disrupted communication 0-did not speak or spoke very little	**10** **7** **4** **1**	
TASK ACHIEVEMENT (10) 3-tasks completed adequately 2-tasks completed almost adequately 1-tasks completed inadequately 0-did not speak or spoke very little	**10** **7** **4** **1**	
TOTAL		

APPENDIX E

INFORMED CONSENT FORM FOR STUDENTS

Consent Form

Dear Student,

I am a student in the MA TEFL 2002 PROGRAM AT Bilkent University. In order to complete my research, I need to make tapes of oral testing sessions. I am asking your permission to tape the session during the speaking midterm in the second midterm week that you will be participating. This taping will not be used to assess your speaking performance. I regard your contribution as a valuable cooperation to my study. All the tapes will be kept in confidential.

If you agree to allow your testing session to be taped, please complete the form below.

Name:
Signature:
Date:

If there are any questions about the study, you may cantact the researcher:
Ece Selva Karslı
MA TEFL Program
Bilkent University

Thank you very much for your cooperation.

Ece Selva Karslı

İzin dilekçesi

Sevgili Öğrenci,

Bilkent Üniversitesi, Yabancı Dil Olarak İngilizce Öğrenimi 2002 programında (MA TEFL) yüksek lisans öğrencisiyim.çalışmamı tamamlayabilmek için sözlü sınav oturumlarının kaydedilmesi gerekiyor. Katılacağınız sözlü sınav oturumunu videoya kaydedebilmek için iznimizi istiyorum. Kayıtlar sizin sınav performansınızı hiçbir şekilde etkilemeyecektir ve benim tarafımdan saklı tutulacaktır. Katılımınız bu çalışma için önemli bir katkıda bulunacaktır.

İsim:
İmza:
Tarih:

Eğer çalışma ile ilgili bir sorunuz olursa, araştırmacı ile iletişim kurabilirsiniz.

Araştırmacı: Ece Selva Karslı
Yabancı Dil Olarak İngilizce Öğretiminde Yüksek Lisans
Programı (MA TEFL Program)
Bilkent Üniversitesi

Katılımınız için teşekkür ederim.

Ece Selva Karslı

APPENDIX F

INTERVIEW QUESTIONS FOR FOUR-BAND SCALE

1. Do you have problems in assessing learners' oral performance?
2. What kind of problems do you have?
3. What do you think about the training session you received before marking?
4. Did you have any problems during the training session?
5. What kind of problems did you have?
6. What do you think about the descriptors in each band in the four-band speaking assessment scale?
7. Are there any terms in the descriptors which are not clear for you?
8. What are these terms?
9. What do you usually do when you cannot decide which band to choose?

APPENDIX G

INTERVIEW QUESTIONS FOR FIVE-BAND SCALE

1. What do you think about the training session you received before marking?
2. Did you have any problems during the training session?
3. What kind of problems did you have?
4. What do you think about the descriptors in each band in the five-band speaking assessment criteria?
5. Are there any terms in the descriptors which are not clear for you?
6. What are these terms?