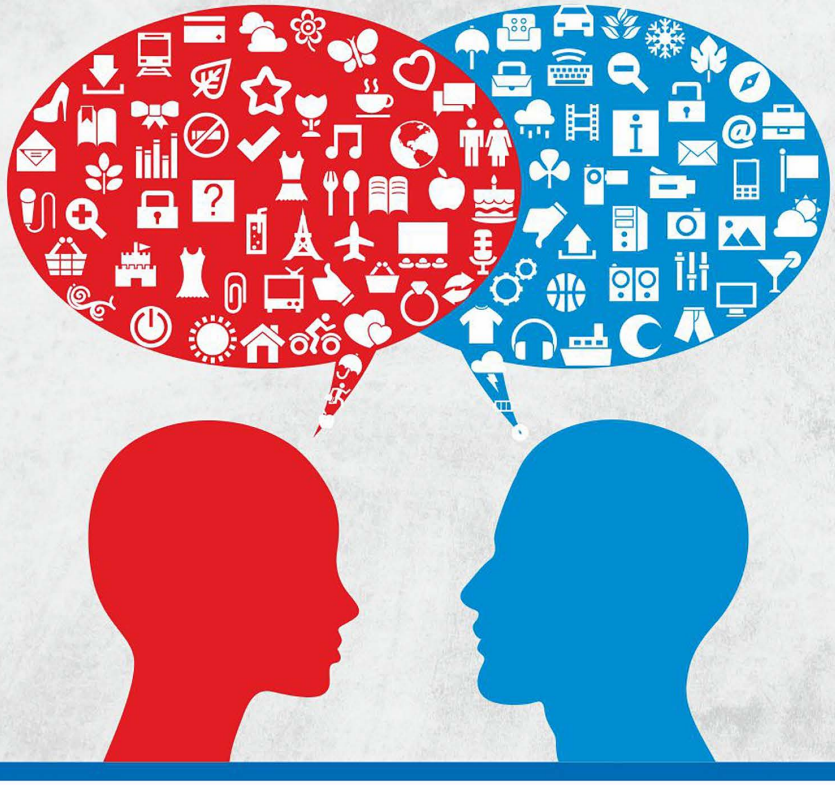
HOLISTIC VS. ANALYTIC ASSESSMENT OF SPEAKING IN EFL - ATTITUDES, SCORES AND THE RATERS



Engin Evrim ÖNEM

astana
yayınları

HOLISTIC vs. ANALYTIC ASSESSMENT of SPEAKING IN EFL - ATTITUDES, SCORES and THE RATERS

Engin Evrim ÖNEM

Ankara
December-2022

HOLISTIC vs. ANALYTIC ASSESSMENT of SPEAKING IN EFL - ATTITUDES, SCORES and THE RATERS
Engin Evrim ÖNEM

© Bu eserin bütün hakları saklıdır.

Sayfa ve Kapak Tasarımı
Astana Tasarım

ISBN
978-625-8045-85-7

YAYIN TARİHİ
20/12/2022

Astana Yayınları
Akademisyen Eğitim Danışmanlık Yayıncılık Hizmetleri A.Ş.
Gökkuşağı Mahallesi 1195 Sok. No: 2/A Çankaya/ ANKARA
Tel: 0 312 230 04 85 – 86 - 87
www.astanayayinlari.com
astanayayinlari@gmail.com

FOREWORD

Speaking is a difficult skill for students to improve and teachers to assess for various reasons. However, although raters are the actual practitioners of these tools, their attitudes toward them are underestimated in most cases. In this respect, this book, adapted from the author's MA thesis, *Instructors' Attitudes towards Assessing Speaking Holistically and Analytically* (submitted to Bilkent University), aims to contribute to closing the gap in the literature by presenting information about relevant background and concepts for holistic and analytic assessment of speaking in ELT as well as a study investigating the attitudes of raters toward them. Furthermore, the study presents a comparison of the exam scores assigned via holistic and analytic assessment tools separately and analyzes them within the light of background variables of the raters who participated in the study along with the discussion and implications of the findings.

Foremost of all, I would like to thank Prof. Dr. Kimberly Trimble, my supervisor, for his support and Assist. Prof. Dr. Deniz Ortaçtepe for her guidance through the preparation of this study. I also owe a great deal to my wife, Dilek, and my daughter İnci İrem for their presence and understanding during this whole tiring but fruitful process. Their presence inspires me greatly...

Engin Evrim ÖNEM

TABLE OF CONTENTS

FOREWORD 3
TABLE OF CONTENTS 4
LIST OF TABLES 5
INTRODUCTION 6
ASSESSMENT IN LANGUAGE TEACHING 8
Assessing Speaking 10
Holistic vs. Analytic Assessment 12
Reliability 17
Raters 19
Factors Affecting Raters 19
THE STUDY ON SPEAKING ASSESSMENT 24
Methodology 27
FINDINGS and DISCUSSION 31
Attitudes toward Holistic and Analytic Assessment 31
Exam Scores Assigned 38
Rater Backgrounds and Exam Scores 39
Conclusion of the Study 42
FINAL THOUGHTS 51
DIRECTIONS FOR THE FUTURE 53
REFERENCES 54
APPENDICES 61

LIST OF TABLES

Table Page

Table 1 - Age groups of the raters 25
Table 2 - Academic major of the raters 25
Table 3 - Years of experience of the raters 26
Table 4 - The results of the holistic assessment questionnaire 31
Table 5 - The results of the analytic assessment questionnaire 32
Table 6 - The themes emerged from the questionnaires 34
Table 7 - Exam scores obtained 38
Table 8 - Exam scores (age of the raters) 39
Table 9 - Exam scores (gender and the holistic assessment tool) 40
Table 10 - Exam scores (gender and the analytic assessment tool) 40
Table 11 - Exam scores (academic major) 41
Table 12 - Exam scores (years of experience) 42

INTRODUCTION

Speaking is an important language skill in learning a foreign language for both students and teachers. As Ur (1996) puts it, a speaker of a language is regarded as someone who has mastered all skills, as speaking contains all other skills. For students, speaking is a very good sign of overall competency in a foreign language since it requires the language user to combine many different linguistic components of a language (morphology, syntax, pragmatics, phonetics, and phonology) at the same in a spontaneous way to convey meaning. One of the reasons most people find speaking extremely difficult is this simultaneity. In this sense, it can easily be accepted that mastery of a language means using the language successfully and the ability to speak in a foreign language is very valuable. For teachers, it is a difficult skill to teach because what the teachers can do about teaching is limited. Besides this limitation of teaching speaking, assessment of speaking is also a problem for many language teachers. Since there are a lot of variables that affect the impression of the teachers as well as the expectancy of testing scores to be accurate (Luoma, 2004), assessing speaking in a foreign language is a real challenge.

Holistic or traditional assessment, focusing on the whole (Brown, 2004), have been used for a long time to assess productive skills such as writing and speaking (Lumley, 2005). In the 1970s, a shift from repetition to production in speaking was seen, and assessing different aspects of speaking at the same time led to a search for assessment tools to provide more precision in testing than the traditional assessment tools (Fulcher, 2003; Luoma, 2004). Because of the precision needed in assessment (Luoma, 2004) and the concerns about assessing different aspects of speaking along with the need for higher rater reliability, there is a shift from the traditional assessment, holistic, to analytic assessment. However, this shift also brings debates about various strengths and weaknesses of analytic and holistic assessment (Brown, 2004; Luoma, 2004; Nakamura, 2004; Weir, 2005). In this sense, analytic and holistic assessments have been compared and analyzed separately in many studies. For instance, while some studies focus on the advantages and disadvantages of both types of assessment (Bachman & Palmer, 1996; 2004; Fulcher, 2003, 2007; Huot, 1990; Luoma, 2004; Nakamura, 2004; Tuan, 2012; Weir, 2005), some focus on

the scores as a result of using both tests (Bacha, 2001; Carr, 2000; Chuang, 2009; Çetin, 2011; Nakamura, 2004) and some focus on the background variables affecting scores (Caban, 2003; Chuang, 2009, 2011) while some focus on the reliability levels of each (Bacha, 2001; Carr, 2000; Chuang, 2009; Çetin, 2011; Nakamura, 2004; Vacc, 1989). There are even some studies focusing on the effect of culture on choosing an analytic or holistic perspective (Monga & John, 2006; Nisbett & Miyamoto, 2005). However, research focusing on the attitudes of teachers toward analytic and holistic assessment is limited (Chuang, 2009; Knoch, 2009). In fact, to the best of the researcher's knowledge, no study focusing on attitudes of teachers towards analytic and holistic assessment of speaking in EFL in Turkey and the scores obtained by holistic and analytic assessment tools have been conducted so far.

As a result, this book has two main aims. First, brief information on related concepts for speaking assessment is presented, and second, a study concerning different types of assessment tools for speaking is presented. In fact, the study presented in the book aims to investigate the attitudes of English teachers towards holistic and analytic assessment of speaking and to reveal if there is a difference in the speaking exam scores assigned by English instructors using holistic and analytic assessment tools. Also, another aim of the study is to find out whether the speaking exam scores obtained via holistic and analytic assessment differ according to the teachers' background (age, gender, academic major, and years of experience).

ASSESSMENT IN LANGUAGE TEACHING

As Brown (2004) puts it, "a test is a method of measuring a person's ability, knowledge or performance in a given domain" (p. 3). Since tests are very common in today's world, testing can be considered "a universal feature of social life" (McNamara, 2000, p. 3). People are tested on different topics every day, such as driving tests or achievement tests, to show that they meet certain criteria or, simply, fit in. When all the social functions of tests are considered, language tests or exams are no exception (Fulcher, 2010). Language tests are instruments that include sets of techniques, procedures, or items to measure a specific or general ability, knowledge, or performance of an individual in an area or areas of language (Brown, 2004). In this sense, tests are closely related to assessment in language learning. As Brown (2004) explains, the assessment of a language is the process of measuring an individual's skills and/or competence in the language. As a result, tests can be considered valuable tools and a subset for assessment, which is an indispensable part of language teaching (Brown, 2004).

There are several reasons why tests and assessments are vital for teaching. First, as Bachman and Palmer (1996) state, tests provide information about students regarding their needs and levels along with feedback on the results of learning and instruction. Administered at the different stages of courses, tests help teachers to get a better view of the current situation. Second, "testing can be used for clarifying instructional objectives and, in some cases, for evaluating the relevance of these objectives and the instructional materials and activities based on them to the language use needs of students following the program of instruction" (Bachman & Palmer, 1996, p. 8). In other words, tests also help institutions to reflect on the teaching/learning process and make revisions if necessary. Among other uses of language tests, although it is rarely discussed, Fulcher (2010) mentions the motivating effects of tests. In fact, according to Fulcher (2010), "when classroom tests were first introduced into schools, an increase in motivation was thought to be one of their major benefits" (p. 1). If students know that they are going to be tested, their attitude towards the course change, and the sooner the test date comes, the more students study. When all these

reasons are considered, it can be seen that tests are important elements of language teaching and learning.

Since testing is an important part of education, the twentieth century sees some major changes in approaches for teaching and assessment in language education, which still prevail today (Brown, 2004). The first approach towards teaching and assessment is "the separate units of language approach". According to Bachman and Cohen (1999), the dominant view of language ability in the 1960s and 1970s is derived from a structuralist linguistic view, which accepts language as being composed of discrete components (grammar, vocabulary) and skills (listening, speaking, reading and writing) to be taught and assessed. As Brown (2004) explains, discrete point tests such as grammar or vocabulary tests are good examples since they are based on the idea that "language can be broken down into its parts and that those parts can be tested successfully" (p. 8).

In response to the structural view, the integrative approach is an alternative and the second major approach seen towards teaching and testing in the 20th century. This approach suggests an indivisible view of language proficiency (Brown, 2004) with a focus on tests that integrated language skills (Bachman & Cohen, 1999; Bachman & Palmer, 1996; Brown, 2004). The best examples of this approach are cloze tests and dictation, where students are expected to integrate different pieces of linguistic knowledge during assessment (Brown, 2004).

As the language teaching and assessment field develops in the 1980s, another approach towards assessment, a competence model based on transformational-generative linguistics, which combines the user's knowledge with performance, appears (Bachman & Cohen, 1999; Brown, 2004). This new paradigm results in a shift in language teaching from the structuralist perspective to a more communicative perspective, in which "a correspondence between language test performance and language use takes place" (Bachman & Palmer, 1996, p. 9). This shift, as Brown (2004) puts it, leads to a quest for authenticity and this is reflected inevitably in language testing as "test designers centered on communicative performance" (Brown, 2004, p. 10). As a result, teaching and assessing productive skills like writing and speaking become as important as receptive skills like reading and listening which are implied in the first two approaches.

Assessing Speaking

Speaking is considered to be the most important skill in learning a foreign language since many people consider it a good sign of overall proficiency. With the emergence of the term "communicative competence" in the early 1970s along with the communicative approach, the emphasis put on speaking in English language teaching (ELT) intensifies (Larsen-Freeman, 2000). However, parallel with the rise of the communicative approach, how to assess speaking becomes a question.

Assessing productive skills, especially speaking, is more difficult than receptive skills for several reasons. First, "from a testing perspective, speaking is special because of its interactive feature" (Luoma, 2004, p. 170). In other words, being an interactive skill makes it harder to assess as it is constantly changing and spontaneous. Weir (2005) emphasizes the spontaneity of speaking and elaborates on the behaviors that are expected in the speech by saying;

> we are no longer interested in testing whether candidates merely know how to assemble sentences in the abstract: we want candidates to perform relevant language tasks and adapt their speech to the circumstances, making decisions under time pressure, implementing them fluently, and making any necessary adjustments as unexpected problems arise. (p. 103)

As known, speaking is a spontaneous production skill, and therefore, during an assessment, "participants have to produce their language according to their resources" (Erlam, 2009, p. 65). In such circumstances, the exam context becomes primary, and meaning along with time pressure affects the context and the performance (Erlam, 2009). This issue raises two questions: First, many studies reveal the relationship between speaking exams and anxiety (see Zeidner, 1998 for a review) and time pressure with anxiety (Hill & Eaton, 1977; Plass & Hill, 1986). As a result, the time pressure may cause an inaccurate evaluation of speaking. Secondly, spontaneous production makes it difficult to control the structures used by the speaker (Erlam, 2009). In other words, assessing spontaneous speaking inhibits selecting and targeting particular aspects of language and results in difficulties in speaking assessment.

The basic nature of speaking, and interaction, is also an issue for assessment. For instance, according to Luoma (2004), since it is always changing, the interactive nature of speaking makes speaking harder to assess. Both the interlocutor and the examinee need to adopt themselves in the course of the interaction. Also, even the nature of the interaction may affect and change the score (Fulcher, 2003). According to Fulcher (2003), the personality or the attitude of the interlocutor may affect the assessment procedure. For example, the friendliness or the unfriendliness of the interlocutor may either have a positive or negative effect on the interaction, which may result in difficulties in terms of reliability in the assessment of speaking.

Another debate in the literature is about the reliability of assessment tools. The term *reliability* refers to the "consistency of measurement of individuals by a test" (McNamara, 2000, p. 136). Inter-rater reliability and intra-rater reliability are the types of reliability related to the raters (Brown, 2004). Inter-rater reliability points to "the extent to which pairs of raters agree" (McNamara, 2000, p. 134), and intra-rater reliability is based on the agreement among scores given by a rater (Brown, 2004). No matter what type of assessment they are, tests are required to have high levels of reliability. Also, raters are the key figures in the assessment of speaking. In fact, they are the necessary but potentially problematic element of assessment (McNamara, 2000) and there are some factors to affect them and cause reliability issues during an assessment. Research has shown that some background variables such as different factors like age, gender, academic major, and years of experience may affect the impression of the instructors about the speaking performance of the language learner (Chuang, 2009). As the main focus of this book, the next section presents two types of assessment for speaking: holistic and analytic as well as issues and studies concerning concepts like reliability and background of the raters.

Holistic vs. Analytic Assessment

In simple terms, holistic assessment, also called traditional assessment (Lumley, 2005), focuses on the whole (Brown, 2004; McNamara, 2000). The most common type of assessment is the traditional, holistic assessment, which has been used for a long time to grade productive skills (Lumley, 2005). Holistic assessment is based on assessing "the performance as a whole" (McNamara, 2000, p. 133). As Brown (2004) summarizes, during a holistic assessment, the rater matches an overall impression with the descriptors to arrive at a score (p. 242). As Alderson, Clapham and Wall (1995) put it, in holistic assessment "examiners are asked not to pay too much attention to any one aspect of a candidate's performance but rather to judge its overall effectiveness" (p. 289). At the end of the holistic assessment, one overall score is assigned either impressionistically, or guided by a rating scale (Fulcher, 2003). Fulcher (2003) goes on to say that "this single score is designed to encapsulate all the features of the sample, representing overall quality" (p. 90). In other words, the impression of the rater depends on the overall quality of the speech sample and is reflected as only a final score at the end of the holistic assessment procedure.

Holistic assessment has both advantages and disadvantages. According to Brown (2004) and Weir (2005), speed is one of them. Since raters do not need to focus on separate components of performance during a holistic assessment, holistic assessment requires less time than analytic assessment. Similarly, one other advantage of holistic assessment is its practicality (Luoma, 2004; Tuan, 2012; Weir, 2005). Because holistic assessment requires the rater to focus on the performance as a whole, there is no need for the attention to be divided among other aspects. As a result, holistic assessment is more practical than analytic assessment. Another advantage of holistic assessment is its flexibility (Huot, 1990). It is known that most raters do not like to be restricted to very specific and limiting sets of criteria (Fulcher, 2010). In turn, holistic assessment enables raters to reflect their impressions of the performance via scores with more freedom. In fact, as Huot (1990) explains, one of the reasons that holistic assessment is widely welcome is that it employs "a rater's full impression of a text without trying to reduce her judgment to a set of recognizable skills" (p. 201). Similarly, Brown (2004) suggests that there is a higher inter-rater reliability (the consistency in the scores given by the same rater) during

a holistic assessment. It is plausible to think that without being limited to a narrow set of criteria, a holistic assessment tool may produce more consistent scoring results.

On the other hand, holistic assessment has also some disadvantages. The most obvious one is using only one score to represent the whole performance. For instance, as Fulcher (2003) puts it;

> it (*holistic assessment*) does not take into account the constructs that make up speaking, but just 'speaking'. And if speaking is made up of constructs, 'speaking' is more like a theory than a construct. A single score may not do justice to the complexity of speaking. (p. 90)

In addition to that, as stated by Fulcher (2003), the final product is composed of different pieces, and ignoring the pieces may lead to inconclusive or misleading results. Similarly, Brown (2004) notes that "one score masks differences across the sub-skills within each score" (p. 242). Also, according to Bachman and Palmer (1996), attempting to represent various components with only one score is not adequate for assessment. Diagnostic inadequacy is another criticism of holistic assessment. As holistic assessment provides little diagnostic information, it limits the positive potential of feedback to students (Brown, 2004; Fulcher & Davidson, 2007; Weigle, 2002). In other words, as the raters' criteria are not explicitly stated, rater scores may not be useful in terms of feedback and washback. Parallel with that, the invisibility of the holistic assessment process is also criticized. As Weigle (2002) suggests, scores of holistic assessment are difficult to interpret because of the differences in the criteria raters had in mind. It is impossible to see the criteria and the process taking place in the minds of the raters, which makes it difficult to control for consistency among raters (Bachman & Palmer, 1996; Fulcher & Davidson, 2007). In fact, Weigle (2002) criticizes this situation as a "tradeoff between high inter-rater reliability at the expense of validity" (p. 114). However, according to Fulcher and Davidson (2007), "within the community of practice, it is precisely the agreement between trained practitioners that is the validity argument" (p. 97). In other words, when raters are trained well, validity may not be an issue for holistic assessment. Yet, since the final score is the only observable outcome of the assessment process, this concerns many.

However, due to the vagueness of holistic assessment, the expectancy that testing scores to be accurate has become an important problem (Luoma, 2004). Similarly, issues about validity and reliability have led to a search for assessing speaking with precision (Lumley, 2005). This leads to a belief that "a precise, empirically based definition of language ability can provide the basis for developing a 'common metric' scale for measuring language abilities in a wide variety of contexts, at all levels, and in many different languages" (Bachman, 1990, p. 5). As a result, the focus turns to analytic assessment tools for speaking in the 1970s (Bachman, 1990), which is based on "assessing each aspect of a performance separately" (McNamara, 2000, p. 131). By using these tools, raters can assess different aspects of speaking in detail, and the process taking place in the raters' minds during assessment can be reflected in the scores easily (Bachman & Palmer, 1996; Brown, 2004). In other words, during a holistic assessment of a productive skill, many variables that the rater tested in his/her mind are reflected as one score but analytic assessment enables outsiders to see the variables and the scores they are assigned separately (Bachman & Palmer, 1996). Analytic assessment is based on "assessing each aspect of a performance separately" (McNamara, 2000, p. 131), which is very different from the holistic assessment. Alderson et al. (1995) explain that analytic assessment is the analysis of a candidate's performance in terms of various components along with descriptors given at different levels for each component. As a result of this multi-component-based analysis, several scores appear at the end of the assessment procedure. The overall score can be calculated by adding up all scores or weighing and valuing the scores differently (Alderson et al., 1995), which depends on the requirements and/or expectations of the institution.

The analytic assessment has some advantages. Foremost among these is the control and consistency of the raters. As Fulcher and Davidson (2007) suggest, test developers can define the extent to be tested by setting up some criteria together. In other words, aspects of the performance to be assessed can be limited and controlled strictly by the test developers. As a result, raters can be directed to pay attention to the aspects of performance, which could otherwise be ignored (Hughes, 2003). In fact, this may also eliminate the uncertainty felt by the examinees before and during the exam (Bachman & Palmer, 1996). In this sense, analytic assessment is a powerful tool to guide the raters.

Also, the scores for each aspect may reflect the actual thoughts and impressions of the rater during the assessment process when analytic assessment tools are used (Bachman & Palmer, 1996; Brown, 2004). In this sense, the analytic assessment makes the assessment process clear to outsiders (Bachman & Palmer, 1996; Fulcher & Davidson, 2007). One other advantage of analytic assessment is the ability to provide a specific set of language abilities to test, differentiate, and weigh components following the expectations (Alderson, Clapham & Wall, 1995; Bachman & Palmer, 1996). This differentiation and weighing refer to the flexibility to change each component's score quotient in the overall score. For example, as analytic assessment focuses on separate aspects, an institution can decide that some aspects should affect the overall score more than others and this makes analytic assessment flexible in terms of expectations. Another advantage of analytic assessment is the feedback it can provide for learners (Fulcher & Davidson, 2007). In this sense, the analytic assessment may be more useful for diagnostic purposes (Luoma, 2004; Weir, 2005). As analytic assessments focus on multiple components of performance, they can provide more details about the performance rather than a simple score (Brown, 2004). Consequently, assessment reports can be shared with examinees so that they have a better understanding of the errors they make.

Despite the advantages of analytic assessment, there are also some disadvantages. First of all, the time analytic assessment takes is a disadvantage. As Hughes (2003) states, the analytic assessment takes more time than holistic no matter how extensively and well the raters are trained. Even preparing clear and relevant criteria suitable for the needs is time-consuming. Secondly, Luoma (2004) highlights the extra cognitive load that comes with analytic assessment. Concentrating on several aspects at the same time may affect raters' focus and may divert them from the overall effect of the assessed work (Hughes, 2003; Luoma, 2004). Therefore, according to Luoma (2004), the raters may do less well during analytic assessment because they need to pay attention to different components at the same time. For instance, while an examinee is responding to a question, the rater using an analytic assessment tool is supposed to divide his attention among different aspects considering various criteria. This may cause an overload in the cognitive capacity and can lead to inaccurate assessments. In fact, this

distraction can undermine the entire assessment procedure. Another criticism of analytic assessment is that it limits the freedom of the raters. As Fulcher (2010) mentions, teachers tend to have negative attitudes toward highly detailed test specifications because such specifications are limiting. As teachers are required to follow specific guidelines during the analytic assessment, they may feel restricted. Similarly, teaching only to meet the criteria can also be another negative outcome of such standards (Fulcher, 2010). Teachers may want to focus on only the aspects represented in the criteria to be assessed and ignore others, which may mean missing pieces in learning. This may have a limiting rather than enriching effect on the implementation of the curriculum (Fulcher, 2010). Another disadvantage of analytic assessment that is often overlooked is the halo effect. Fulcher (2010) explains the halo effect as "a phenomenon where the act of making one judgment colors all subsequent judgments" (p. 209). In fact, Weir (2005) raises a question about the halo effect;

> the possibility exists that the rating of one criterion might have a knock-on effect in the rating of the next. If a major preoccupation of a marker is with grammar, and the candidate exhibits a poor performance on this criterion, are the marks awarded in respect of other criteria contaminated by the grammar mark? (p. 188-189)

In other words, when analytic assessment tools are used, the score of one component of the performance may have either a positive or negative effect on the proceeding score of the component. In addition to positive or negative effects, it is suggested that

> in language testing we commonly find that if a rating is made on one scale it is carried over to others. The effect is the creation of a flat profile, even if a learner is, in fact, more proficient in some areas than others. (Fulcher, 2010, p .209)

As a result, the analytic assessment may have a higher degree of halo effect on the components of the performance.

To sum up, none of these assessment types are superior to each other in every aspect, and deciding which type to use in assessing any kind of performance is up to the user and/or the institution (Bachman

& Palmer, 1996; Fulcher, 2010; Luoma, 2004) since both holistic and analytic assessment have various strengths and weaknesses.

Reliability

Reliability is an important issue regardless of the type of assessment tool used. As McNamara (2000) explains, reliability is the "consistency of measurement of individuals by a test" (p. 136). In other words, it is the consistency of the test to yield the same scores. Although there are different types of reliability such as student-related, test administration, or test reliability, there are two rater-related types of reliability: inter-rater reliability and intra-rater reliability (Brown, 2004). Inter-rater reliability refers "to the extent to which pairs of raters agree" (McNamara, 2000, p. 134) while intra-rater reliability refers to the agreement among scores given by a single rater (Brown, 2004). Regardless of the type of assessment such as holistic or analytic, tests are expected to display high levels of reliability. However, studies report that changing the rating scale type, namely, from holistic to analytic or vice versa affects scores and the whole reliability issue related to the type of assessment is controversial in the literature. In fact, it is possible to find studies reporting a change in the reliability level depending on the assessment tool being used. For example, as Carr (2000) reports on the findings, both the interpretation of a section of a test and total test scores can differ depending on the assessment tool being used. Similarly, in the study conducted by Barkaoui (2010), findings indicate that the rating scale type, either holistic or analytic, has a large effect on the scores. In a study by Song and Caruso (1996), a statistically significant difference in the scores obtained by holistic assessment is found but no such difference is found in the analytic scores, which suggests a higher reliability level in analytic assessment. Similarly, Nakamura (2004) reports a higher inter-rater reliability level for analytic scoring than holistic scoring. For instance, Chuang (2009) compares the scores obtained via analytic and holistic assessment of the same instructors separately and reports no significant difference between the scores. Similarly, Vacc (1989) and Bacha (2001) find a strong relationship between the scores obtained via analytic and holistic assessment and report high inter-rater reliability levels in the tests. In contrast, some studies report a higher level of reliability in favor of the holistic assessment. For instance, Barkaoui (2007) compares the holistic and analytic scores of 24 EFL essays and reports

that holistic scoring shows higher inter-rater reliability than analytic scoring. Similarly, O'Loughlin (1994) makes a comparison regarding the holistic and analytic scores and finds a higher inter-rater reliability level in holistic assessment than in analytic assessment. Yet, O'Loughlin (1994) suggests that even though holistic assessment seems to be more reliable, it is less valid because holistic assessment may be masking the differences among the raters' scores, unlike analytic assessment.

As would be expected, there are also some studies reporting high levels of reliability when the scores obtained via holistic and analytic assessment tools are compared. For example, in a study, Bacha (2001) focuses on the scores of two sets of essays assigned by different raters holistically and analytically. The results indicate that the raters achieve high levels of inter-reliability and intra-reliability using both holistic and analytic assessment tools and there is no statistical difference between the reliability levels for both types of assessment. In his study, Çetin (2011) analyzes analytic and holistic scoring for writing assessment in three different ways; holistic-holistic, analytic-holistic, and analytic-analytic. Çetin (2011) reports that high inter-rater reliability is seen in holistic-holistic and analytic-analytic comparison but when holistic and analytic scores are compared, there is a lower level of inter-rater reliability. In her study, Chuang (2009) asks raters to assign scores for oral performance by using both holistic and analytic assessment tools and checks the scores to see whether there is a difference in the scores. Results show high inter-rater reliability and no statistically significant differences in inter-rater reliability between the scores obtained via both types of assessment (Chuang, 2009).

To avoid the issues of reliability in terms of assessment tools, Luoma (2004) suggests that the needs and/or the expectations of the institution should have a key role in choosing the relevant assessment type. In other words, either holistic or analytic assessment should be chosen depending on the things to be assessed. In fact, there are also some studies suggesting that using or combining holistic and analytic assessment tools at the same time may lead to better assessment performances (Harsch & Martin, 2012; Jin, Mak & Zu, 2012). Yet, given the controversies in the literature about the advantages and disadvantages of both assessment types as well as the scores and the rater reliability issues, it would not be surprising to expect variation

among the attitudes of language instructors towards analytic and holistic assessment. Yet, more research is required.

Raters

Regardless of the tool of assessment, the actual practitioners of assessment are generally neglected. The term, raters, is frequently used interchangeably with assessors or scorers in the literature. Yet, the *rater* is the most commonly used term and this study uses it to describe the language instructor who is using an assessment tool to evaluate and score the performance of a student during an exam.

Raters are the key people in assessment "who judge performances in productive tests of speaking and writing, using an agreed rating procedure and criteria in so doing" (McNamara, 2000, p. 136). They are the bridge between the assessment tool with valid and reliable assessment. However, as McNamara (2000) states, including the raters during the assessment process is necessary as well as problematic. McNamara (2000) emphasizes that rating is a subjective process and is depended on the rater, especially for the assessment of productive skills. Each rater's impression and expectancy for what makes a good performance varies (Luoma, 2004). In fact, McNamara (2000) goes on to say "the rating given to a candidate is a reflection, not only of the quality of the performance but of the qualities as a rater of the person who has judged it" (p. 37). Although training the raters seems to work to some extent, total elimination of differences in the scores seems impossible (Wang, 2010; Weigle, 1994), and "not all human elements can be compromised in the evaluation process" (Vanniarajan, 2006, p. 290). As a solution to this problem, Wang (2010) suggests creating *rater files* that include information for every rater in an institution about their tendency in scoring so that they can be selected more appropriately for the assessment task. Yet, it is quite plausible to find factors affecting raters differently.

Factors Affecting Raters

Although there are several factors affecting the raters such as "their mother tongue, age, gender, educational background, research areas, knowledge about ESL learning and oral ability development, personal character, experience as a rater, whether they have received any training to be raters, etc.," (Wang, 2010, p. 109), this book and the

study presented in focuses only on years of experience, academic major, age and gender of raters because these factors seem to be the most effective in terms of raters and the scores they assign.

Age is closely related to years of experience in teaching. However, studies comparing scores of teachers with different age groups in ELT are limited. For example, Chuang (2009) compares scores assigned to the speaking performance of EFL learners by four groups of teachers of different ages. The results show a statistically significant difference among groups with the youngest raters in the study (21-30) scoring the lowest while the oldest group (50+) scoring the highest. In other words, the raters of the age group over 50 are the most lenient in terms of scoring and the younger teachers become the more severe scores they seem to assign.

The literature on the effect of the gender of the raters on the scores they assign is mixed. On one hand, some studies report that male raters score higher than female raters. For instance, Locke (as cited in Chuang, 2011) analyzes the scores of male and female raters on the oral performance of EFL learners. The results show a statistically significant difference between the scores, with male raters scoring higher than female raters. Similarly, Porter (1991) reports that among a range of variables, including the personality of the participants and their degree of acquaintanceship, the only variable that has a significant effect on the students' oral performance is gender, with the scores of male raters higher than female raters. On the other hand, some studies report that female raters assign higher scores than male raters. For example, in a study by Gholami et al. (2011), both male and female raters are asked to interview learners and assess their oral proficiency separately. Although the inter-rater reliability levels in both interviews rated by male and female raters are high, students receive higher scores from the female raters and the difference is statistically significant (Gholami et al, 2011). Similarly, O'Sullivan and Porter (1996) focus on the interview scores of ESL learners assigned by male and female raters and report that female raters' scores are statistically significantly higher in the study. Yet, some studies report findings that reveal no difference between the scores given by the male and female raters. For instance, in a study by O'Loughlin (2002) making a comparison between male and female raters' scores in oral proficiency testing in IELTS, it is seen that gender does not have a significant effect on the scores. Similarly, Chuang

(2011) reports no statistically significant difference in terms of the scores assigned by the male and female raters. In other words, the gender of the raters does not have a direct effect on the scores the raters assign in some studies. Interestingly, in some studies, results indicate that learners with the same gender as the rater receive a higher score. For example, Buckingham (as cited in Chuang, 2011) reveals that male students get higher scores when being interviewed by a male rater or vice versa. To sum up, regardless of all the different results in the literature, the gender of the rater seems to be a variable that may affect the scores assigned.

Academic major or the department graduated from is another background variable for the raters using different types of assessment tools. In a study, Chuang (2009) analyzes the scores of raters with different academic majors (literature/linguistics, TESOL/ESL/ EFL, others). Results show that a statistically significant difference is seen among the scores of the raters whose academic major is literature/linguistics and TESOL/ ESL/EFL. The scores of the raters with literature/linguistics majors are the lowest while the scores of the raters with teaching majors are the highest and the results were consistent within the groups. However, the difference between these two groups and linguistics/literature is not significant (Chuang, 2009). This result may suggest that because the expectations of the literature/linguistics major raters are high, they tended to score low. On the other hand, raters with teaching majors may be considered more tolerant of errors than others and this tendency is reflected in their scores. In another study, Chuang (2011) asks teachers with different majors such as literature, linguistics, and teaching to evaluate the speaking performances of 75 students of EFL students using a holistic assessment tool. The results reveal that academic background has a certain degree of impact on the test scores assigned by the raters. Deeper analyses show that the scores of the raters with linguistics or literature majors are significantly lower than the raters with TESOL backgrounds and other major backgrounds.

Years of experience is an expected factor for speaking assessment and there are some studies in the literature focusing on the effect of years of experience on assessment. For instance, Chuang (2009) conducts a study about English teachers' scoring performance for speaking and focuses on the effect of teaching experience and other

backgrounds. The results indicate no statistically significant difference between years of experience, which means the scores of the teachers for speaking assessment do not differ in terms of years of experience in that EFL case. However, in another study by Song and Caruso (1996), it is found that raters with more years of experience tend to be more lenient when using holistic assessment tools. In a more recent study, similar results are seen. Huang and Jun (2015) focus on raters determining the native-likeness of the speech samples and compare three groups of raters' (inexperienced, experienced, and advanced) scoring performance for speech production. The findings show that inexperienced raters are stricter in their ratings than both of the other groups (Huang & Jun, 2015). As these findings suggest, there might be a tendency in the raters to be more tolerant as they become more experienced and it seems like raters' years of experience may affect the scores they assign.

Although language instructors are the actual user of assessment tools, there are only a few studies in the literature on the attitudes of instructors toward holistic and analytic assessment. Of the limited studies, Knoch (2009) focuses on teachers' perceptions of a scoring scale with less specific descriptors and a scale with more detailed descriptors. The results show that raters prefer to use the more detailed, analytic scale. The main reason for that is the raters in that study believe that the more detailed scale helps them to focus more on the details rather than an overall impression. Interestingly, most of the raters consider the more detailed scale as "minimally more time consuming" (Knoch, 2009, p. 298) and some even consider it faster. In another study, Barkaoui (2010) compares novice and experienced teachers and their perceptions of holistic and analytic scales. He finds that novice and experienced teachers have different perceptions of the use of the scales. Barkaoui (2010) summarizes the findings as follows;

> first, novice teachers show a shift from a focus on specific linguistic features (e.g., syntax, lexis, spelling) with the analytic scale to a focus on rating language overall with the holistic scale. Second, both groups tended to refer more often to linguistic appropriacy with the analytic scale. Finally, the novices referred more frequently to text organization when rating the essays analytically, suggesting that the analytic scale drew their attention to

> this aspect of writing as well as linguistic appropriacy. (p. 64)

Barkaoui (2010) also reports that novice teachers tend to refer to the rating scales more frequently than experienced raters. In other words, novice teachers check whether they are consistent with the scales more often than experienced teachers do. As Barkaoui (2010) concludes, "although the two groups differed in terms of several strategies, the differences across scales are more noticeable" (p. 65). Maybe, the more experienced a teacher becomes, the more freedom s/he wants during assessment since s/he may have already created a set of invisible criteria in their mind. Similarly, less experienced teachers may lack such criteria and may be dependent on outer sources during the assessment. Yet, as limited studies suggest, there is a difference in terms of the teachers' perceptions of using holistic and analytic assessment tools.

THE STUDY ON SPEAKING ASSESSMENT

As mentioned above, the aim of the book is both to summarize the relevant literature on speaking assessment and present the findings of a study to address a gap in the literature. Therefore, in this section, the methodology of the study is presented since the study aims to find out the attitudes of raters for holistic and analytics assessment of speaking, whether there are differences or not in the scores obtained by using them along with the effect of rater background on the scores. Therefore, in this section, first, the research design, setting and participants of the study are introduced. Secondly, the instruments used to collect data as well as methods of data collection, which include both quantitative and qualitative data collection methods, and methods of data analysis are presented.

The study includes both quantitative and qualitative data to triangulate the results and obtain a better understanding of the phenomena. Therefore, as Creswell (2003) and Brown and Rodgers (2002) state, this study can be considered to have a mixed method design that incorporates both quantitative and qualitative data collection methods rather than subscribing to only one way (Creswell, 2003, p. 12). As a result, both multiple-choice questions and open-ended items are included in this study as proposed by Creswell (2003). For instance, an attitude questionnaire prepared by the researcher and holistic and analytic assessment tools previously used at a state university is used to obtain quantitative data about the attitudes of the instructors toward speaking exam assessment. Two open-ended items which are included at the end of the attitude questionnaire to gain deeper insights into the attitudes of the instructors are the main source of qualitative data in this study.

The study took place at a state university in Turkey for five months. Since both holistic and analytic assessment tools are used previously at the host institute, instructors who have experience using both assessment types of speaking are included only as participants. As a result, twenty-four full-time language instructors participate in the study voluntarily. The background variables of the instructors who participate in this study include gender, age, years of experience they have in teaching English, and academic majors (the department they graduated from).

As for the gender of the instructors, fourteen of them are male and ten of them are female. As for the ages of the instructors, their ages range from 26 to 46+. The highest number of participants for the age groups is in the 31-35 years old group. The frequency for the age groups of the instructors is given in Table 1.

Table 1 - Age groups of the raters

Age Groups	N
26-30	2
31-35	8
36-40	5
41-45	2
46+	7

For the academic major of the instructors, there is a variety of departments including ELT, Literature, and Linguistics. The instructors who have graduated from ELT departments form the largest group in terms of academic majors. The distribution of the academic major for the instructors is given in Table 2.

Table 2 - Academic major of the raters

Academic Major	N
ELT	13
Literature	9
Linguistics	2

The instructors have varying years of experience in teaching English and they are grouped into five categories since all the instructors who participate in the study have more than at least five years of experience. Instructors with 11-15 years of experience compose the largest group in terms of years of experience. The distribution of the years of experience can be seen in Table 3.

Table 3 - Years of experience of the raters

Years of Experience	N
6-10	4
11-15	9
16-20	5
20+	6

To sum up, although the sample size is limited to twenty-four instructors, there is a variety of background variables in terms of age, gender, years of experience, and academic major.

Both quantitative and qualitative data collection tools are used in the study to obtain data related to the attitudes of the instructors along with speaking exam scores. For instance, after reviewing the literature on analytic and holistic assessment (Bachman & Palmer, 1996; Brown, 2004; Fulcher, 2003; Hughes, 2003; Huot, 1990; Luoma, 2004; Weigle 2002; Weir, 2005), the researcher creates a questionnaire depending on the advantages and disadvantages of both types of assessment. The four-point Likert questionnaire consists of 20 multiple choice items which are designed to reveal the attitudes of the raters toward the holistic and analytic assessment of speaking concerning the advantages such as ease of use or effectiveness of each type of assessment. The questionnaire is composed of two sets of questions for the holistic and the analytic assessment of speaking. Ten items (items # 2, # 3, # 5, # 8, # 9, # 13, # 15, # 16, # 19, and # 20) are used to reveal the attitudes of the instructors towards holistic assessment of speaking and items # 3 and # 20 are reverse coded. As for the items related to analytic assessment, items # 1, # 4, # 6, # 7, # 10, # 11, # 12, # 14, # 17, and # 18 are used and items # 12 and # 14 are reverse coded. When analyzing the table, it should be remembered that the highest value (4) is assigned to the response, *Strongly Agree*, while the lowest score (1) is assigned to *Strongly Disagree*. Reverse-coded items are re-coded in SPSS.

A test of internal consistency is run to see whether the questionnaire has high internal consistency. The *Cronbach's alpha* for each set of questions is calculated as .93 for holistic assessment and .92 for analytic assessment. As a result of consulting with an expert and high numbers of internal consistency, the questionnaire is accepted as reliable (see Appendix A).

For assessing the speaking performances of students, holistic and analytic tools assessment tools are used in the study. Both tools are developed and validated by the Testing Office of the aforementioned state university, which is responsible for test administration and evaluations of the scores. Since both tools are used previously for assessing speaking exams given at the host institute, the tools are considered reliable and valid for this study as well.

The holistic assessment tool includes a basic set of descriptors for raters to follow during assessment and a range of scores for the performance of the student, among which the raters are asked to assign only one score they feel is appropriate for the overall performance (see Appendix B). The analytic assessment tool, on the other hand, includes different components of speaking such as pronunciation, grammar, vocabulary, fluency, and comprehension for raters to assess the learners' speaking performance. It has a more detailed set of descriptors for each component of speaking separately. It also includes a range of scores for each component of speaking for the raters to choose to assign and the final score is calculated by adding up the scores assigned for each component (see Appendix C).

Along with the multiple-choice items included in the attitude questionnaire, two open-ended items designed for the instructors to write the reasons for their positive or negative attitudes for both types of assessment are also included at the end of the questionnaire. The items are semi-structured since they require instructors to choose between "positive/negative attitude" and then elaborate on the reason for their choice by providing extra information about their attitudes for each type of assessment (see Appendix B).

Methodology

The researcher obtained the holistic and analytic assessment tools which are used previously at the host institute before the data collection process as well as the *Oral Exam Procedure Booklet* from the *Testing Office* of the very same institute (see Appendix D for the guidelines), which includes directions to be paid attention during speaking exams. These directions include information on how to conduct a speaking exam, suggestions, and sample questions or tasks. After being informed about the study, ten students studying English at the host university accept to participate in the study. Since each student is placed in the pre-

intermediate classes after the proficiency exam given previously at the beginning of the Fall term, they are all regarded as students with similar levels of English. These students are given a speaking exam individually by the researcher in two days and the exams take place at the office of the researcher at the university. The speaking exams take approximately five minutes and consist of a warm-up section (approx. one minute) and a question section (approx. four minutes). All the questions asked during the speaking exam are chosen from the *Oral Exam Procedure Booklet* (see Appendix E for sample questions). The exams start as normal conversations (warm-up section) and then the researcher asks questions from the *Oral Exam Procedure Booklet*. Since the level of the students is pre-intermediate, the researcher assists the students when they need help by rephrasing the questions, changing the topic, or asking follow-up questions. Although there is not a constant number of questions asked, students are asked approximately seven to ten questions. During the speaking assessment, all of the questions and students' oral responses are recorded using a sound recorder by the researcher as an audio file (see Appendix F for a transcribed sample of the exam).

Due to a large number of students and feasibility issues, assessing the speaking performances of the learners is stopped long before this study. Therefore, although only the instructors who are familiar with using both assessment tools participate in this study, they are nevertheless trained to use both analytic and holistic assessment tools separately by the researcher. Each training session takes place in the conference hall and each session lasts for an hour. The instructors are trained to use the holistic assessment tool first. During the training session, the researcher explains the rubric to the instructors and gives information about the assessment process. The researcher holds another training session for the analytic assessment tool a month later. Similar to the training session for the holistic assessment, the rubric is explained and information about the assessment process is given. All of the instructors participate in the assessment process and attend both training sessions.

The data is collected as follows: Ten randomly selected prep class students studying English are recorded separately during the aforementioned speaking exam in English given by the researcher. After obtaining the holistic and analytic assessment tools from the institution,

twenty-four language instructors from the same institution who have experience in using both types of assessment tools are trained to use the holistic assessment tool at first. At the end of the training session, all of the instructors listen to the audio recordings of the speaking exam given by the researcher together at the same time only once without any interruptions at the conference hall. While listening to the recordings, they are asked to assess the speaking performances of the students by using the holistic assessment tool given by the researcher. The process for the holistic assessment takes approximately fifty minutes and the tools and the scores are collected at the end of the process. Later, a similar procedure is followed for the analytic assessment of the speaking exams. However, to minimize the raters' recalling of the speaking exams and the scores they assign, the analytic assessment of the speaking exams is held a month later. The instructors listen to the audio recordings of the same speaking exam together at the same time only once without any interruptions again and use the analytic assessment tool to assess the speaking performances of the students provided by the researcher. The process takes approximately fifty minutes and at the end of the process, the tools and the scores given are collected again.

Following the week of the analytic assessment, the instructors are administered the attitude questionnaire prepared by the researcher, which is composed of 20 four-point Likert Scale items and two open-ended questions to reveal their attitudes towards holistic and analytic assessment. They are asked to choose the option they feel reflects their attitudes towards holistic and analytic assessment at best and provide the reasons for their choices in as much detail as possible through the open-ended items. The instructors fill in the questionnaires in their own spare time and the completion of the questionnaire takes a week. At the end of the data collection procedure, both quantitative and qualitative data are obtained. The quantitative data consists of the scores obtained from the attitude questionnaire and the speaking exam scores obtained from holistic and analytic assessment tools. The qualitative data include the answers to the open-ended questions for the instructors' reasons for their attitudes towards different types of assessment.

To reveal the attitudes of the instructors towards holistic and analytic assessment of speaking, the quantitative data obtained from the attitude questionnaire are analyzed by descriptive statistics while

the qualitative data are analyzed thematically. Means and standard deviations for each item are calculated and presented. The thematic analysis is based on the frequency of the positive and negative elements that are seen in the responses of the participants and similar elements are grouped under the same themes. To see whether the speaking exam scores assigned by using different types of assessment differed or not, *a paired samples t-test* is used. To find out whether the speaking exam scores assigned by using holistic and analytic assessment differed according to the background variables of the instructors, *an independent samples t-test* and *one-way ANOVA* is used.

FINDINGS and DISCUSSION

In this section, the attitudes of the raters toward holistic and analytic assessment of speaking are presented separately and this is done concerning the four-point Likert Scale item scores of the questionnaire. Then, the analysis of the qualitative data obtained via the open-ended items at the end of the questionnaire is explained.

Attitudes toward Holistic and Analytic Assessment

The questionnaire related to the attitudes toward holistic assessment of speaking consists of ten items: # 2, # 3, # 5, # 8, # 9, # 13, # 15, # 16, # 19, and # 20. Items # 3 and # 20 are reverse coded. The results of the descriptive statistics analysis are presented in Table 4.

Table 4 – The results of the holistic assessment questionnaire

Item #*	$\overline{x}$	SD
3	3.37	.711
13	3.21	.658
5	3.21	.779
19	3.08	.776
9	3.00	.978
2	3.00	.780
15	2.96	.806
20	2.92	.776
8	2.88	.850
16	2.42	.881

N=24, * in descending order of means

During the analysis of the table, it should be kept in mind that four points are assigned to *Strongly Agree*, 3 are assigned to *Agree,* 2 are assigned to *Disagree* while the lowest score (1) is assigned to *Strongly Disagree*. Reverse-coded items are re-coded in SPSS.

The descriptive analysis of the results of the attitudes for holistic assessment shows that item # 3, *Holistic assessment is time-consuming*, receives the highest mean score ($\overline{x}$ = 3.37, SD = .711). It should be kept in mind that the item is reverse-coded and it is re-coded. In this sense, the participants have the most positive attitude towards this item, implying that holistic assessment is not time-consuming. The second most positive attitude toward holistic assessment is seen in two items:

13, *Grading with holistic assessment is easier than using analytic assessment* ($\overline{x}$ = 3.21, SD = .658) and # 5, *Holistic assessment enables the rater to score more freely* ($\overline{x}$ = 3.21, SD = .779). Following these items, item # 19, *Holistic assessment is better than analytic assessment in terms of practicality* ($\overline{x}$ = 3.08, SD = .776), item # 9, S*peaking performance should be assessed as holistically* ($\overline{x}$ = 3.00, SD = .978) and item # 2, *With holistic assessment, I feel the scores I assign reflect my true impressions* ($\overline{x}$= 3.00, SD = .780) receive positive scores from the instructors. Of the final four items, which receive lower scores than the previous items, there is a slight difference in terms of their means. For instance, the attitudes of the instructors are slightly lower for # 15, *I feel more confident and comfortable when assessing performance with a holistic assessment tool* ($\overline{x}$= 2.96, SD = .806), # 20, *At the end of the holistic assessment, the final score is higher or lower than I expected* ($\overline{x}$ = 2.92, SD = .776) and # 8, *Holistic assessment allows the rater to reflect his/her impressions of the performance better* ($\overline{x}$= 2.88, SD = .850). Item # 16, *Better feedback for students can be provided with holistic assessment* ($\overline{x}$= 2.42, SD = .881) get the least positive score.

The questions related to the analytic assessment consist of ten items as well. To reveal raters' attitudes towards analytic assessment for speaking, items # 1, # 4, # 6, # 7, # 10, # 11, # 12, # 14, # 17, and # 18 are used. Items # 12 and # 14 are reverse coded. The results can be seen in Table 5.

Table 5 – The results of the analytic assessment questionnaire

Item #*	$\overline{x}$	**SD**
7	2.92	.881
4	2.42	.776
17	2.38	.711
18	2.33	.816
10	2.33	1.049
6	1.96	.806
14	1.87	.797
1	1.83	.637
11	1.83	.702
12	1.79	.779

N=24, * in descending order of means

The descriptive analysis of the results of the analytic assessment questionnaire shows that item # 7, *Analytic assessment can provide better feedback for students*, has the highest mean score ($\overline{x}$= 2.92, SD = .881). Following that item, item # 4, *Using an analytic assessment tool makes me feel more confident and comfortable* ($\overline{x}$= 2.42, SD = .776), # 17, *A better impression of the performer can be obtained via analytic assessment* ($\overline{x}$= 2.38, SD = .711), # 18, *Speaking performance should be assessed analytically* ($\overline{x}$= 2.33, SD = .816), and # 10, *I feel the scores I assign reflect my true impressions when using analytic assessment tools* ($\overline{x}$= 2.33, SD = 1.049) have lower and yet similar scores. The final five items receive very similar scores and they are the lowest items in the analytic assessment questionnaire. These items are: # 6, *In terms practicality, analytic assessment is better than holistic assessment* ($\overline{x}$= 1.96, SD = .806), # 14, *The final score ends up often higher or lower than I expected at the end of the analytic assessment* ($\overline{x}$= 1.87, SD = .797), # 1, *Using analytic assessment is easier than using holistic assessment* ($\overline{x}$= 1.83, SD = .637), # 11, *Analytic assessment allows the rater to grade the performance with more freedom* ($\overline{x}$= 1.83, SD = .702) and # 12, A*nalytic assessment takes a long time* ($\overline{x}$= 1.79, SD = .779).

Along with the 20 Likert-scale items on the questionnaire, two semi-structured open-ended items are included to get a more detailed view of the raters' attitudes towards holistic and analytic assessment. Of these two open-ended items, the first one, *As a language instructor, I have a positive/negative attitude towards analytic assessment because…*, focuses on the raters' attitudes towards analytic assessment while the second item, *As a language instructor, I have a positive/negative attitude towards holistic assessment because…*, focuses on the attitudes for holistic assessment. The two open-ended items are analyzed thematically concerning raters' attitudes toward different types of assessment. In other words, the responses to the two open-ended items are analyzed by grouping the similar themes that emerged from the answers of the instructors and are presented separately in this section. The positive and negative elements that emerged from the open-ended items are given in Table 6.

Table 6 - The themes emerged from the questionnaires

	Positive elements	**Negative elements**
Holistic assessment tool	- the practicality of holistic assessment - the idea for assessing speaking performance holistically - true reflections of the raters' impressions with holistic scoring	- the subjectivity of holistic assessment - the vagueness the rating process - the necessity of training instructors to use holistic assessment
Analytic assessment tool	- the feedback opportunities - the reliability of the scores - the comfort	- the time needed to complete analytic assessment - assessing speaking as a whole, not piece by piece - the inability of reflecting the true impressions of the rater because using analytic assessment - dividing attention into separate aspects

Several themes are observed in the raters' responses to the open-ended questions. The most common theme to emerge in the attitudes of the instructors is *the practicality of holistic assessment.* Eleven out of fourteen of the instructors who have positive attitudes towards holistic mention that holistic assessment is more practical. One of the instructors suggests that it is because of its convenience for their (assessment) purposes. The instructor goes on to explain that due to the high number of students taking the exams in a limited time, it is easier to assign only one overall score. One other instructor highlights that it is difficult to follow many criteria and the student at the same time and that's why holistic assessment is more practical. As a result, it can be

inferred that the instructors with positive attitudes towards holistic assessment believe that the practicality of the assessment tool is an important issue.

Another similar theme seen is *the idea of assessing speaking performance holistically*. One of the instructors supports the holistic assessment of speaking by saying "speaking performance should be considered as a whole and so it should not be graded from many aspects", highlighting the notion that the whole is bigger than the pieces. In a similar vein, another instructor claims that holistic assessment is more realistic and practical than analytic assessment because people take assessments as a whole in real life, and "so should speaking be".

Another theme observed in the open-ended items is the ability to give *true reflections of the raters' impressions with holistic scoring*. One of the instructors writes that s/he feels more confident about the score s/he assigned to a student's performance when using holistic assessment tools because s/he could reflect his/her real impressions more easily with one score. Another instructor emphasizes the freedom holistic assessment creates and explains "I feel that I have given the most accurate score by using holistic scoring since you are more free".

On the other hand, several negative different themes emerge in the raters' criticisms of the holistic assessment. The first theme seen is *the subjectivity of holistic assessment*. As one of the instructors writes, "the score obtained by the holistic assessment seems to be more personal, which may lead to reliability issues". According to another instructor, "since scoring is done more freely in holistic assessment, the scores can be misleading sometimes". One instructor who shares similar ideas goes on to say that many aspects of speaking could be missed during holistic assessment due to the absence of a controlled and guided assessment tool like analytic scoring tools. As a result, s/he mentions that s/he does not feel that the scores obtained via holistic assessment reflected his/her true impressions. These comments suggest that clearer guidelines are needed to ensure the reliability of the scores.

Another theme observed has a similar concern about *the vagueness* of *the rating process*. One of the instructors criticizes the holistic assessment as a very vague type of assessment, by which no clear and visible assessment process could be achieved. Another

instructor writes "it (holistic scoring) does not provide concrete reasons for the grade I assign to my students" and emphasizes the issue of subjectivity and vagueness of the holistic assessment process one more time. One final critical theme of holistic assessment is *the necessity of training instructors to use holistic assessment*. One of the instructors suggests that unless teachers are trained in how to use the holistic criteria, the scores could not be considered reliable. Another instructor also points to rater training and states that the raters should be trained in advance to raise the level of reliability, which, otherwise, would be an issue for holistic assessment. As a whole, these negative elements are critical of holistic assessment for being vague and subjective.

Just like the themes that emerged for holistic assessment, different but related themes are also seen in the themes for analytic assessment. The most common positive element of analytic assessment is *the feedback opportunities* that it provides. Seven out of ten instructors with a positive attitude towards analytic assessment agree that analytic assessment provides feedback opportunities for the students. Some of them mention that this could help students to see the specific speaking performance points they need to improve using analytic assessment.

The second theme observed is *the reliability of the scores* obtained by analytic assessment. For instance, instructors seem to agree that analytic assessment tools produce scores with higher reliability. Although there is not any specific information provided by the instructors, the term "reliability" is assumed to be referring to both inter and intra-reliability of the raters. For instance, one of them states that since "you have a more detailed series of guidelines, the reliability of the scores assigned would be higher". Another instructor points to the need for guidelines to reach more reliable assessment results. As a result, the instructors with positive attitudes towards analytic assessment seem to pay attention to the reliability issues in assessment.

The final positive element in the raters' responses towards analytic assessment is *the comfort of analytic assessment*. The instructors who explain their reasons for their positive attitude in the open-ended items state that they are comfortable with the guided and controlled nature of the assessment tool. As noted by one instructor, "having specific categories in our hands while evaluating my students' performance makes me feel very comfortable". In other words, it can be

said that the instructors who have a positive attitude toward analytic assessment seem to believe that having clear-cut guidelines is very useful. In sum, those who view analytic assessment positively seem to do so because of the feedback it provides to students, the reliability of the scores, and their comfort with specific scoring guidelines.

On the other hand, the most common criticism that emerges in the responses is the quantity of *time needed to complete an analytic assessment.* Nearly all of the instructors who indicate negative attitudes towards analytic assessment emphasize this theme clearly. For instance, one of them writes "it (analytic assessment) takes a lot more time than it should" and another one explains "focusing on different aspects takes a lot of time and is not practical". Another instructor writes that analytic assessment is very time-consuming because it needs to focus on different aspects at the same time, which is a tiring type of assessment.

The second theme that emerged is the emphasis on *assessing speaking as a whole, not piece by piece*. For example, one of the instructors writes "the quality of pieces does not make any sense for me if the whole body does not function/work", signaling his/her positive attitude toward holistic assessment. Another instructor warns that during the analytic assessment, "teachers begin to ignore the general picture while analyzing the smaller parts" and as a result, the reliability of the assessment is lowered, quite the opposite from what would have been expected from analytic assessment.

Another theme seen in the open-ended items is *the inability of reflecting the true impressions of the rater because of using analytic assessment*. One of the instructors mentions that "analytic assessment makes scoring complicated and this prevents the rater to reflect his/her impression of true performance". A second instructor explains that analytic assessment is misleading in terms of the final score. It can be assumed that the instructors who suggest these ideas are referring to adding up the scores of separate sections to reach a final score at the end of the analytic assessment. They seem to believe that as a result, the final score might be higher or lower than what is anticipated.

Similarly, *dividing attention into separate aspects* is the final theme that emerges. One of the instructors writes "it (analytic assessment) takes a long time and I have to divide my attention between separate

categories of the performance. It is tiring and energy consuming". Similarly, another instructor suggests "it is difficult to follow the criteria and the student at the same time", which would lead to missing important parts of the performance. These ideas suggest that the extra cognitive load to keep up with the performance while assessing different aspects of a performance can fatigue the rater and may result in a decrease in the rating performance. To sum up, the critiques of analytic assessment seem to revolve around time consumption, the difficulty in reflecting the raters' real impressions of the performance, and the difficulty in assessing different parts simultaneously.

Exam Scores Assigned

To answer the second research question, whether speaking exam scores assigned by holistic assessment differ from scores using the analytic assessment tool, means of the speaking exam scores obtained via holistic and analytic assessment tools are compared using *a paired samples t-test*. The scores given by each instructor using the holistic and analytic assessment tool can be seen in Table 7.

Table 7 - Exam scores assigned (general)

	Holistic Score		**Analytic Score**	
Student #	$\overline{x}$	**SD**	$\overline{x}$	**SD**
1	46.46	13.85	51.96	13.80
2	64.88	10.06	65.75	12.30
3	75.92	14.74	76.21	13.91
4	51.25	13.89	52.21	14.11
5	58.46	15.13	57.38	14.85
6	44.29	12.06	45.25	9.94
7	69.42	11.40	68.13	13.48
8	43.04	14.25	48.42	12.82
9	50.96	13.61	47.33	13.14
10	63.38	14.86	60.00	13.01

The results do not yield a statistically significant difference between the scores obtained by the holistic assessment tool ($\overline{x}$= 56.80, SD = 8.44) and analytic assessment tool ($\overline{x}$= 57.26, SD = 9.76, t (23) = .190, p > .05) and there is only a slight difference between the mean scores of the scores assigned using different assessment tools. As a result, it could be said that no matter which type of assessment tool is

used to assess speaking performance at the sample speaking exams in this study, the scores assigned are nearly the same.

Rater Backgrounds and Exam Scores

As for the third research question, whether the speaking scores assigned by the instructors using holistic and analytic assessment tools differed according to the raters' background, the means of the scores assigned for the speaking exams via holistic and analytic assessment tools are compared following the age, gender, academic major and years of experience of the instructors.

The data obtained from the holistic and analytic attitude questionnaire are compared according to age groups by employing *One-Way ANOVA*. However, although there are six groups of age (21-25/26-30/31-35/36-40/41-45/46+), there is not a participant in the first group of 21-25. The results of *One-Way ANOVA* are presented in Table 8.

Table 8 - Exam scores (age of the raters)

		SS	***df***	**MS**	***F***	***p***
Holistic Scores	Between Groups	327.727	4	81.932	1.187	.348
	Within Groups	1311.362	19	69.019		
	Total	1639.090	23			
Analytic Scores	Between Groups	192.308	4	48.077	.457	.766
	Within Groups	1998.323	19	105.175		
	Total	2190.636	23			

As can be seen in the table, results show no statistically significant difference among different age groups for the scores they assign by holistic assessment [F (4, 19) = 1.187, p >.05] or analytic assessment tools [F (4, 19) = .457, p >.05]. The results may suggest that in the present study, the instructors of different ages assign similar scores for the sample speaking exams when using both assessment tools.

The data obtained from the holistic and analytic tools for speaking exams are compared according to gender by employing *an independent samples t-test*. The results of the *t*-test regarding gender and the

speaking exam scores obtained by the holistic assessment tool are presented in Table 9.

Table 9 - Exam scores (gender and the holistic assessment tool)

	Male		Female					
Holistic Score	$\bar{x}$	**SD**	$\bar{x}$	**SD**	**N**	**r**	***t***	***df***
	54.26	6.31	60.36	10.01	24	.36*	1.83*	22

*p >.05

The results show that for the scores assigned by using holistic assessment, there is not a statistically significant difference between male participants (x = 54.26, SD =6.31) and female participants ($\bar{x}$= 60.36, SD = 10.01), t (22) = 1.831, p > .05). It can be inferred from the results that male and female instructors assign similar scores using holistic the assessment tool.

The results of the statistical analyses for gender and the speaking exam scores obtained by the analytic assessment tool can be seen in Table10.

Table 10 - Exam scores (gender and the analytic assessment tool)

	Male		Female					
Analytic Score	$\bar{x}$	**SD**	$\bar{x}$	**SD**	**N**	**r**	***t***	***df***
	58.35	9.06	55.74	10.97	24	.13*	.638*	22

*p >.05

Similarly, the results show that for the scores obtained by using the analytic assessment tool, there is not a statistically significant difference between male participants ($\bar{x}$= 58.35, SD = 9.06) and female participants ($\bar{x}$ = 55.74, SD = 10.97), t (22) = .638, p > .05). Therefore, it can be said that male and female participants assign similar scores using the analytic assessment tool as well.

The academic majors of the participants are ELT, Literature, and Linguistics departments. The data obtained from the holistic and analytic assessment tools for speaking exam scores are compared according to the academic major of the participants by employing *One Way ANOVA*. The results of the *One Way ANOVA* are given in Table 11.

Table 11 - Exam scores (academic major)

		SS	***df***	**MS**	***F***	***p***
Holistic Scores	Between Groups	312.950	2	156.475	2.478	.108
	Within Groups	1326.140	21	63.150		
	Total	1639.090	23			
Analytic Scores	Between Groups	30.064	2	15.032	.148	.865
	Within Groups	2160.572	21	102.884		
	Total	2190.636	23			

Results show that there is not a statistically significant difference between participants' academic majors and the scores they assign by using holistic [F (2, 21) = 2.478, p >.05] or analytic assessment tool [F (2, 21) = .148, p >.05]. As a result, it may be considered that the instructors with different academic majors assign similar scores using holistic assessment and analytic assessment in the present study.

Although there are five groups for years of experience (1-5/6-10/11-15/16-20/20+), there is not a participant in the first group of 1-5, and the speaking exam scores assigned using holistic and analytic assessment tools are compared following years of experiences of the instructors by employing *One Way ANOVA*. The results are presented in Table 12.

Table 12 - Exam scores (years of experience)

		SS	*df*	MS	*F*	*p*
Holistic Scores	Between Groups	605.483	3	201.828	3.905	.024
	Within Groups	1033.606	20	51.680		
	Total	1639.090	23			
Analytic Scores	Between Groups	505.166	3	168.389	1.998	.147
	Within Groups	1685.470	20	84.274		
	Total	2190.636	23			

Results show that there is a statistically significant difference between participants' years of experience and the scores they assign using the holistic assessment tool [F (3, 20) = 3.905, p <.05] and the effect size was large η^2 = .37 between groups. The result of a post hoc LSD test indicates that only the differences between 6-10 and 11-15 and 16-20 and 20+ are not statistically significant. However, when the groups with different years of experience are analyzed more closer, the less experienced instructors seem to score higher. For example, instructors with 6-10 years of experience have the highest mean scores ($\overline{x}$ = 62.05, SD = 7.79) and is followed by the 11-15 year experience group ($\overline{x}$ = 61.13, SD = 8.24). However, the scores assigned by the other two groups of 16-20 ($\overline{x}$ = 51.28, SD = 4.22) and 20+ ($\overline{x}$= 51.42, SD = 6.89) are the lowest. On the other hand, there is not a statistically significant difference in the scores obtained via the analytic assessment tool and the years of experience of the instructors [F (3, 20) = 1.998, p >.05].

Conclusion of the Study

The findings of the study in this book are presented in this chapter separately for each research question. In the first section, the findings related to the raters' attitudes toward holistic and analytic assessment of speaking are discussed. In the second section, the findings concerning the speaking exam scores obtained via holistic and analytic assessment tools are presented. In the third section, the findings concerning the speaking exam scores and the background variables of the instructors are explored.

The analysis of the results for both the four-point and semi-structured open-ended items in the attitude questionnaire shows that instructors in this study have different attitudes toward holistic and analytic assessment. First, except for one item, the mean scores for all the items of the holistic assessment are higher than the mean scores for nine out of ten items of the analytic assessment. In other words, the results of the questionnaire show that the instructors who participated in the study show a more positive attitude towards holistic assessment of speaking.

Secondly, when the items in the questionnaire and the positive and negative themes that emerged from the responses to the open-ended items are analyzed, several overarching themes are seen, mostly in favor of the holistic assessment. The most common support for holistic assessment and criticism for analytic assessment seen is the issue of *true reflections of the impressions of the rater*. For instance, item # 2, *With holistic assessment, I feel the scores I assign reflect my true impressions, # 8, Holistic assessment allows the rater to reflect his/her impressions of the performance better*, # 10, *I feel the scores I assign reflect my true impressions when using analytic assessment tools*, # 14, *The final score often ends up often higher or lower than I expected at the end of the analytic assessment*, # 17, *The rater can reflect his/her impression of the performance better via analytic assessment* and a reverse coded item, # 20, *At the end of the holistic assessment, the final score is often higher or lower than I expected* reflect the participants' attitudes towards holistic and analytic assessment. As different aspects of speaking performance are focused on simultaneously during analytic assessment and the final score is the result of a combination of different scores for the different aspects, some of the instructors complained that after adding the scores for different aspects, the final score might end up different from the scores they think. For example, one of the instructors writes that since s/he is supposed to give only one score during the holistic assessment, the score is a real reflection of his/her impression. However, when s/he is asked to use the analytic assessment tool, he has to give separate scores and then add them. The final score, he says, is generally higher or lower than he had anticipated. Another theme observed is *practicality*. Four items in the questionnaire such as item #1, *Using analytic assessment is easier than using holistic assessment, # 6, In terms of practicality, analytic assessment is better than*

holistic assessment, # 19, Holistic assessment is better than analytic assessment in terms of practicality and # 13, *Grading with holistic assessment tools is easier than using analytic assessments* are highly related to this theme. The most common reason given for *practicality* is the ease of holistic assessment provided for the assessment of the whole. When the workload of instructors is considered, practicality is a big issue. Therefore, it is quite understandable that raters in the present study show a positive attitude towards this aspect of holistic assessment in terms of *practicality*. *The time needed for assessment* is another theme that was seen in the attitudes of the participants concerning holistic and analytic assessment. This theme matches with item # 3*, Holistic assessment is time-consuming* and # 12, *Analytic assessment takes a long time.* In fact, a reverse-coded item, # 3 has the highest mean score, which means that there is a strong agreement among the participants with the statement and # 12 receives the least positive score. One of the instructors writes that trying to score different aspects at the same time is a hard thing to do. Another one writes, "instead of one score, you try to give five or six scores, which makes it harder and harder". As a result, the participants seem to pay attention to the time assessment takes and that is another reason for them to have a positive attitude toward holistic assessment. *The freedom allowed by the assessment tool* is another theme that emerges from the responses of the participants, which turns out to be in favor of the holistic assessment. This theme matches with item # 5, *Holistic assessment enables the rater to score more freely* and # 11, *Analytic assessment allows the rater to grade performance with more freedom*. One of the instructors explains that s/he does not like being controlled and trying to make his scores fit into the criteria and that is why s/he prefers holistic assessment. When the mean scores for each item for this theme are looked at, the difference in the mean score can lead people to consider that the participants thought holistic assessment allows them to score with more freedom. *Assessing speaking holistically* is another theme supported by items # 9, *Speaking performance should be assessed holistically* and # 18, *Speaking performance should be assessed analytically*. Nine out of ten instructors who give this as one of the reasons for their positive attitudes towards holistic assessment explain that the whole is bigger than the pieces and speaking performance should be assessed as a whole. As a result, it can be said that the themes

also seem to support the positive attitudes of the participants toward holistic assessment.

Yet, the most common theme in favor of analytic assessment seen is the *feedback opportunities it created*. Not surprisingly, item # 16, *Better feedback for students can be provided with holistic assessment* receive the least positive attitude score, indicating that on average participants *disagreed* with the statement while the instructors seem to *strongly agree* with item # 7, *Analytic assessment can provide better feedback for students*. The instructors consider analytic assessment as a more useful type of assessment to provide feedback for students to improve their skills. For instance, one of them writes "at the end of the analytic assessment, I know what my students need to do to develop themselves". Finally, *the comfort analytic assessment created* is another theme that emerges in favor of the analytic assessment. Item # 15, *I feel more confident and comfortable when assessing performance with a holistic assessment tool* has a higher mean score than item # 4, *Using an analytic assessment tool makes me feel more confident and comfortable*. According to one of the instructors, having guidelines makes his/her job easier since following guidelines is easy. Yet, this element is only seen in the responses of the participants to the open-ended items in favor of analytic assessment and is not seen for holistic assessment.

Third, some themes highlight different issues than those mentioned in the items of the attitude questionnaire. For example, the most common criticism of holistic assessment is *subjectivity of holistic assessment*. The six instructors who criticize holistic assessment for being subjective suggest that stricter guidelines are required for more reliable scores. For example, one of them writes "when I use the analytic rubric, it guides me, and the scores I assign become more reliable". *The reliability of the scores* obtained by analytic assessment is another point made by instructors. The instructors seem to believe that analytic assessment provides more reliable scores than those provided by holistic assessment. For example, one instructor explains that "to reach more reliable scores, guides are needed and analytic scoring provides this". Parallel with that, *explaining the rating process* is another common criticism of holistic assessment. The four instructors write about their concern for the *vagueness of the rating process* when holistic assessment is used. In fact, one of them criticizes it as a tool lacking concrete guidelines, which, inevitably, leads to unreliable scores. Another

criticism of holistic assessment is related to rater training. The instructors point to *the necessity of training instructors to use holistic assessment* so that the scores could be regarded as reliable. Otherwise, the scores would be very subjective and vague. One final theme, *a division in attention*, that comes up only in the responses of the participants is about the analytic assessment and the cognitive load it brings. In fact, one instructor writes that s/he could not follow the student when s/he has to focus on different aspects of his/her performance and this could lead to a failure in the assessment because he has to deal with many aspects simultaneously.

The findings related to the attitudes of the instructors towards holistic and analytic assessment reveal some aspects of speaking assessment that instructors pay attention to. First of all, it seems like the ease of use and time are among the major issues the instructors consider when it comes to assessment. They seem to consider these two as critical and have a tendency to favor the type of assessment that enables them to assess easily and quickly. This might be one of the reasons for the positive attitude toward holistic assessment. Secondly, the instructors in the study seem to believe that speaking should be assessed as a whole rather than from different aspects or components. As mentioned above, they tend to consider speaking skill as the final performance which includes all of the pieces already. This might also be related to their expectation for an assessment tool to be practical and easy to use and those expectations are met by holistic assessment. For instance, since holistic assessment enables the instructors to assess the performance as a whole with one score only, this might be another reason for the emergence of a positive attitude towards holistic assessment. Another possible reason for the instructors to have positive attitudes towards holistic assessment is related to reflecting true impressions of the rater about the performance. Since analytic assessment focuses on the components of performance separately, the final score obtained might end up different from what is anticipated and the instructor might feel that s/he has not reflected his/her true reflections about the performance. On the other hand, since only one score is assigned at the end of the holistic assessment, it is easier to reflect one's impressions. Finally, the *subjectivity of the assessment* tended to be the strongest argument for the attitudes of the instructors towards analytic assessment over holistic assessment in this study. This

seems to be an issue with the holistic assessment as one final score to cover all components is thought to lead to unexplainable scores.

When these findings are considered together, it is seen that the attitudes of the instructors seem to be parallel with the literature in terms of the advantages and disadvantages of both holistic and analytic assessment. For instance, as mentioned by Brown (2004), Luoma (2004), and Weir (2005), *practicality* was an advantage of the holistic assessment. In fact, as Luoma (2004) highlights, this might alone be one of the important reasons for instructors to choose holistic assessment over analytic assessment. Similarly, Huot (1990) highlights *flexibility* or *more freedom* as an advantage of holistic assessment, and this is seen as another advantage the instructors in this study mentioned. On the other hand, as Brown (2004) and Weigle (2002) point out, holistic assessment offers no feedback or had no washback effect for students but analytic assessment does. This is the strongest criticism for holistic assessment and is put forward for the analytic assessment by the instructors. The raters' other criticism of the holistic assessment is related to the impossibility to monitor the assessment process. According to Bachman and Palmer (1996), during the holistic assessment, monitoring the assessment procedure is very difficult and the instructors in the study seem to have noticed it as well. Also, as Brown (2004) and Luoma (2004) emphasizes, the subjectivity of the scores is one of the most common criticisms seen both in the literature and the responses of the instructors in this study. As for the analytic assessment, control and consistency provided (Hughes, 2003) and more feedback opportunities for students (Brown, 2004; Fulcher & Davidson, 2007) are among the reasons the instructors support analytic assessment of speaking. Also, feeling more comfortable because of having a more detailed scoring tool (Knoch, 2009) is another reason for the positive attitude toward analytic assessment. Yet, being time-consuming (Brown, 2004; Hughes, 2003*),* requiring more attention and/or a division in the attention (Brown, 2004; Luoma, 2004), and limiting the raters (Fulcher, 2010) are among the criticism put forward by the participants in this study for analytic assessment.

The findings suggest that regardless of the tool used to assess speaking performance in this study, the scores assigned by the instructors do not show any statistically significant difference. This finding contradicts Carr (2000) and Barkaoui (2010), who report that

the type of assessment tool used affects the scores. However, the findings of this study are parallel with those of Chuang (2009), who finds no statistically significant difference among the scores obtained via holistic and analytic assessment tools. There might be several underlying reasons for this result. For one, the examinees' levels of English were all pre-intermediate. As a result, it is plausible to suggest that the homogeneity of the speaking samples might affect the results. Another reason might be related to the instructors, which might be called the decision-making behaviors of the instructors (Barkaoui, 2007). Although the instructors are asked to assess the speaking exams analytically, they might have ignored the components of the tools, assign an overall score and then divide them into components of the assessment tool.

To answer the third research question, the mean scores of the speaking exam scores obtained via holistic and analytic assessment tools are compared with the age, gender, academic major, and years of experience of the participants.

The assigned speaking scores are compared by age groups of the participants by employing One Way ANOVA and it is found the scores do not differ in terms of age groups. The findings are different from the findings of Chuang (2009), which state that younger teachers are more severe in terms of scoring while older teachers are more lenient. The reason for the nonsignificant difference in this study might have been caused by the low number of participants. For example, if the sample size is bigger, with more participants of different ages, this might result in a difference in the findings for the difference in scores. Also, as mentioned above, the homogeneity of the examinees' level of English might have affected the findings for age and scores.

As for the comparison of the speaking exam scores by the gender of the participants, *an independent samples t-test* is used and it is seen that the scores do not show any statistically significant difference according to the gender of the instructors. This finding adds a little more information to the body of literature since there are mixed results reported in the literature. For instance, although Locke (as cited in Chuang, 2011) and Porter (1991) report that male teachers scored higher than female teachers, Gholami et al. (2011) and O'Sullivan and Porter (1996) report the opposite findings in their studies. However,

the findings in this study seem to be parallel with those reported by O'Loughlin (2002) and Chuang (2011), which find no statistically significant difference in terms of the scores of male and female teachers, which clearly show that the gender of the instructors does not lead to a difference in the scores.

The participants in this study have various academic majors, including ELT, Literature, and Linguistics departments. Comparison of the holistic and analytic speaking scores with raters' academic majors employing *One-Way ANOVA* shows no statistically significant difference. This finding is different from the findings reported in the literature. For instance, Chuang (2009) and Chuang (2011) report that the scores assigned by the raters with a literature/linguistics major are lower than the scores of the raters with a teaching major. The reason for such a difference might be explained by the fact that since the participants are working as English instructors, no matter what their academic major was, this might override their educational background and leads them to be more teaching-oriented. As a result, regardless of their academic major, they become teachers of English in the end and became more tolerant of the errors of the students, which is reflected in the scores they assign. However, further research is required to prove this.

The comparison of the speaking exam scores obtained from holistic and analytic assessment tools with years of experience of the instructors by employing *One Way ANOVA* shows interesting results. For instance, although a statistically significant difference is not seen in terms of analytic assessment, there is a statistically significant difference between participants' years of experience and the scores they assign using the holistic assessment tool. Specifically, the less experienced instructors are more lenient (scored higher) than more experienced instructors.

The findings can be considered from two aspects: more or less experienced instructors and holistic and analytic assessment tools. The finding in this study that less experienced raters are more lenient seems to be different from the studies reported in the literature. For instance, although Chuang (2009) indicates no statistically significant difference between years of experience, Song and Caruso (1996) find that raters with more years of experience tend to be more lenient when using holistic assessment tools. Similarly, Huang and Jun (2015) report that inexperienced raters are stricter in their ratings than the other group

(Huang & Jun, 2015). The reason for such a finding in this study can be related to the fact that less experienced raters may be less familiar with the assessment tool and/or assessing speaking. If longitudinal research is conducted, this finding might be confirmed or rejected.

The fact that such a difference in scores is seen only in scores assigned by the holistic assessment tool can be explained by the stricter structure of the analytic assessment tool. Since analytic assessment is more guided, regardless of the years of experience, instructors in this study seem to assign similar scores but when they are more free and without stricter guidelines, the less experienced instructors seem to score more freely. As a result, such a difference in the scores is seen.

FINAL THOUGHTS

This book has two main aims: to highlight points concerning the assessment of speaking via holistic and analytic assessment tools and to present a relevant study. Similarly, the study in this book has two main objectives: revealing the attitudes of language instructors towards holistic and analytic assessment and determining whether scores for speaking assessment obtained by holistic and analytic assessment differ or not. The findings primarily suggest that instructors in this study have different attitudes towards holistic and analytic assessment, which are based on their ideas about the assessment of speaking. For example, the instructors who have a positive attitude toward the holistic assessment seem to reject piece by piece assessment of speaking and vice versa. Yet, the findings also show that instructors have a more positive attitude toward holistic assessment. Second, the findings reveal that there is not a statistically significant difference between the scores using the holistic and analytic assessment instruments.

Even though more data is required, yet, within light of these findings, it can be argued that after being trained well on both types of assessment, instructors may be allowed to use one type of assessment of their choice for actual assessment practices. Since the scores are close to each other in this study and instructors display varying attitudes towards holistic and analytic assessment considering the advantages and disadvantages of both types of assessment, allowing them to choose the tool to use may help them to perform better during the assessment. As instructors have a preference for the type of assessment tool they use, using the assessment they prefer might lead to more satisfaction for the instructors and the institution. Also, at the pedagogical level, the findings of this study may help teachers of ELT in different parts of Turkey by presenting a snapshot at a state university in Turkey about the current status of the holistic and analytic assessment debate. Then, they could have a better idea of the current status in terms of assessing speaking as well as differences and similarities between scores obtained by different types of assessment and could reflect the findings to their testing processes. Also, by receiving insights from both sides of the analytic vs. holistic debate, it may help administrators and testing professionals in other institutions to review and/or plan their testing aims.

Another pedagogical implication of the study is related to using different types of assessment tools simultaneously. Though this is not the focus of this study, there are studies suggesting using holistic and analytic assessment simultaneously during assessment (Harsch & Martin, 2012; Jin, Mak & Zu, 2012), which would require at least two raters. It is, then, quite plausible to obtain more reliable and valid scores for assessment.

This book can contribute to the literature in three dimensions. First, the findings of the study can help to fill in the gap in the literature by revealing the attitudes of EFL instructors towards assessing speaking both analytically and holistically. Secondly, since a debate about analytic and holistic assessment is present in the literature, the findings of the study can reveal the opinions and the position of the actual users of different assessment tools in terms of the debate. Finally, the findings related to the scores may contribute to the literature by revealing whether scores obtained by holistic and analytic assessment differ or not,

DIRECTIONS FOR THE FUTURE

Firstly, related to the pedagogical implications presented in the previous section, another study may focus on whether raters' preferences regarding the assessment tools affect the scores they assign. In other words, allowing instructors to choose between using the holistic or analytic assessment tool and comparing the results of an assessment procedure to see whether a difference occurred or not can be studied. Secondly, the only instructors who had experience in using both types of assessment tools are included in the study. In other words, novice instructors do not participate in the study. A future study may include novice instructors to see whether their attitudes and/or scores would differ from more experienced instructors. Third, because of time and feasibility constraints, the instructors are only asked to assess the sample speaking exams silently. However, in future studies, they can be asked to think aloud during the process of scoring. Thus, the process taking place in their minds during assessment can be followed for further analysis and this could give a deeper understanding of their choice in terms of types of assessment. Finally, the level of all sample speaking exams is pre-intermediate. As a result, including speaking performances from different levels of English might have a different effect on the results. Another study could include more students with different levels of English and more instructors to see whether different levels of English make difference in terms of speaking exam scores obtained by holistic and analytic assessment tools.

REFERENCES

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

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring can tell us? *System, 29*, 371-383.

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

Bachman, L. F., & Cohen, A. D. (1999). Language testing - SLA interfaces: An update. In L. F. Bachman, & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 1-31). Cambridge, England: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford: Oxford University Press.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12*(2), 86-107.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*, 54-74.

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice, 18*(3), 279-293.

Blake, R. J. (2011). Current trends in online language learning. *Annual Review of Applied Linguistics, 31*, 19-35.

Brown, H. D. (2001). *Teaching by principles: An integrative approach to language pedagogy.* New York: Pearson Education.

Brown, H. D. (2004). *Language assessment - Principles and classroom practices.* NY, USA: Pearson Education.

Brown, J. D., & Rodgers, T. S. (2002). *Doing second language research.* Oxford: Oxford University Press.

Caban, H. L. (2003). Rater group bias in the speaking assessment of four L1 Japanese ESL students. *Second Language Studies, 21*(2), 1-44.

Carr, N. T. (2000). A comparison of the effects of analytic and holistic rating scale types in the context of composition tests. *Issues in Applied Linguistics, 11*(2), 207-241.

Chi, E. (2001). Comparing holistic and analytic scoring for performance assessment with many-facet Rasch model. *Journal of Applied Measurement, 2*(4), 379-388.

Chuang, Y. Y. (2009). Foreign language speaking assessment: Taiwanese college English teachers' scoring performance in the holistic and analytic rating methods. *The Asian EFL Journal Quarterly, 11*(1), 152-175.

Chuang, Y. Y. (2011). How teachers' background differences affect their rating in EFL oral proficiency assessment. *Studies in English Language and Literature*(28), 37-55.

Creswell, J. W. (2003). *Research design - Qualitative, quantitative, and mixed methods approaches.* Thousand Oaks, CA: SAGE Publications.

Çetin, Y. (2011). Reliability of raters for writing assessment: analytic - holistic, analytic - analytic, holistic – holistic. *Mustafa Kemal University Journal of Social Sciences Institute, 8*(16), 471-486.

Erlam, R. (2009). The elicited oral imitation test as a measure of implicit knowledge. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp, & H. Reinders, *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 66-93). NY, USA: Multilingual Matters.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*, 208-238.

Fulcher, G. (2003). *Testing second language speaking.* NY, USA: Taylor & Francis.

Fulcher, G. (2010). *Practical language testing.* London: Hodder Education.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment - An advance resource book.* Oxford: Routledge.

Gholami, J., Sadeghi, K., & Nozad, S. (2011). Interviewers' gender and interview topic in oral exams. *Theory and Practice in Language Studies, 1*(10), 1394-1399.

Harsch, C., & Martin, G. (2012). Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice, 20*(3), 281-307.

Hill, K. T., & Eaton, W. O. (1977). The interaction of test anxiety and success-failure experiences in determining children's arithmetic performance. *Developmental Psychology, 13*(3), 205-211.

Huang, B. H., & Jun, S. A. (2015). Age matters and so may raters-Rater differences in the assessment of foreign accents. *Studies in Second Language Acquisition*, 1-28. doi:10.1017/S0272263114000576

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge, UK: Cambridge University Press.

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication, 41*(2), 201-213.

Jin, T., Mak, B., & Zhou, P. (2012). Confidence scoring of speaking performance: How does fuzziness become exact? *Language Testing, 29*(1), 43-65.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275-304.

Larsen-Freeman, D. (2000). *Techniques and principles in language teaching.* Oxford: Oxford University Press.

Lightbown, P. M., & Spada, N. (2013). *How languages are learned* (4th ed.). Oxford, UK: Oxford University Press.

Lumley, T. (2005). *Assessing second language writing: The raters' perspective.* Frankfurt: Peter Lang.

Luoma, S. (2004). *Assessing speaking.* Cambridge, UK: Cambridge University Press.

Lynch, B. K. (1997). In search of the ethical test. *Language Testing, 14*(3), 315-327.

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

Monga, A. B., & John, D. R. (2006). Cultural differences in brand extension evaluation: The influnce of analytic versus holistic thinking. *Journal of Consumer Research, 33*, 529-536.

Murray, D. E., & Christison, M. A. (2011). *What English Language Teachers need to know? Understanding Learning* (Vol. 1). Oxford: Routledge.

Nakamura, Y. (2004). A comparison of holistic and analytic scoring methods in the assessment of writing. *3rd annual JALT Pan-SIG Conference* (pp. 45-52). Japan: Tokyo Keizai University.

Nisbett, R. E., & Miyamoto, Y. (2005). The influence of culture: Holistic versus analytic perception. *Trends in Cognitive Science, 9*(10), 467-473.

Nunan, D. (1992). *Research methods in language teaching.* Melbourne: Cambridge Language Teaching Library.

O'Loughlin, K. (1994). The assessment of writing by English and ESL teachers. *Australian Review of Applied Linguistics, 17*(1), 23-44.

O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing, 19*(2), 169-192.

Oruç, N. (1999). Evaluating the reliability of two grading systems for writing assessment at Anadolu University prepoartory school. *Unpublished MA thesis*. Ankara, Turkey: Bilkent University.

O'Sullivan, B., & Porter, D. (1996). Speech style, gender, and oral proficiency interview performance. *Paper presented at the Anual Meting of the Southeast Asian Ministers of Education Organization Regional Language Center Seminar.* Singapore. Retrieved December 25th, 2014, from http://files.eric.ed.gov/fulltext/ED403744.pdf

Owens, J. R. (1996). *Language development.* USA: Allyn & Bacon.

Önem, E. E. (2015). Raters' Attitudes towards Assessing Speaking Holistically and Analytically. *Unpublished M.A. Thesis*. Ankara: Bilkent University.

Plass, J. A., & Hill, K. T. (1986). Children's achievement staretegies and test performance: The role of time pressure, evaluation anxiety and sex. *Developmental Psychology, 22*(1), 31-36.

Porter, D. (1991). Affective factors in language testing. In J. C. Alderson, & B. North (Eds.), *Language testing in the 1990s* (pp. 32-40). London, Modern English Publications in association with The British Council: Macmillan.

Song, B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking ESL students? *Journal of Second Language Writing, 5*(2), 163-182.

Sowden, C. (2007). Culture and the 'good teacher' in the English language classroom. *ELT Journal, 61*(4), 304-310.

Tuan, L. T. (2012). Teaching and assessing speaking performance through analytic scoring approach. *Theory and Practice in Language Studies, 2*(4), 673-679.

Ur, P. (1996). *A course in language teaching: Practice and theory.* Cambridge: Cambridge University Press.

Vacc, N. N. (1989). Writing evaluation: Examining four teachers' holistic and analytic scores. *The Elementary School Journal, 90*(1), 87-95.

Vanniarajan, S. M. (2006). Human factors in holistic assessment of writing a heuristic (pilot) study. *Proceedings of the 11th Conference of Pan-Pacific Association of Applied Linguistics*, (pp. 290-306).

Wang, B. (2010). On rater agreement and rater training. *English Language Teaching, 3*(1), 108-112.

Weigle, S. C. (1994). Using FACETS to model rater training effects. *Paper presented at the 16th Annual Language Testing Research Colloquium.* Washington DC.

Weigle, S. C. (2002). *Assessing Writing.* Cambridge: Cambridge University Press.

Weir, C. J. (2005). *Language testing and validation - An evidence based approach.* New York: Palgrave Macmillan.

Zeidner, M. (1998). *Test anxiety - The state of art.* New York, USA: Plenum Press.

APPENDICES

Appendix A - The Attitude Questionnaire

Please circle the correct information for you

Age:21-25/26-30/31-35/36-40/41-45/46+ **Years of Experience**:1-5/6-10/11-15/16-20/20+
Gender: Male / Female **Academic Major**: ELT/Literature/Linguistics

INSTRUCTIONS

Below is a questionnaire about language instructors' attitudes towards analytic and holistic assessment of speaking. Please read all items carefully and choose the option suits best to you. At the end of the questionnaire, there are two open ended items. Please complete the items to reflect your attitude towards analytic and/or holistic assessment in your own words. You may use the back side of the paper if you have other comments. All your answers will be stored anonymously. Thank you for your cooperation.

	Strongly disagree	Disagree	Agree	Strongly agree
Using analytic assessment is easier than using holistic assessment.				
With holistic assessment, I feel the scores I assign reflect my true impressions.				
Holistic assessment is time consuming.				
Using an analytic assessment tool makes me feel more confident and comfortable.				
Holistic assessment enables the rater to score more freely.				
In terms of practicality, analytic assessment is better than holistic assessment.				
Analytic assessment can provide better feedback for students.				
Holistic assessment allows the rater to reflect his/her impressions of the performance better.				
Speaking performance should be assessed as a holistically.				
I feel the scores I assign reflect my true impressions when using analytic assessment tools.				
Analytic assessment allows the rater to grade performance with more freedom.				
Analytic assessment takes a long time.				
Grading with holistic assessment tools is easier than using analytic assessments.				
The final score often ends up often higher or lower than I expected at the end of the analytic assessment.				
I feel more confident and comfortable when assessing performance with a holistic assessment tool.				
Better feedback for students can be provided with holistic assessment.				
The rater can reflect his/her impression of the performance better via analytic assessment.				
Speaking performance should be assessed analytically.				
Holistic assessment is better than analytic assessment in terms of practicality.				
At the end of the holistic assessment, the final score is often higher or lower than I expected.				

As a language instructor, I have a **positive / negative** attitude towards **analytic assessment** because _____
_____.

As a language instructor, I have a **positive / negative** attitude towards **holistic assessment** because _____
_____.

Appendix B – The Holistic Assessment Tool

HOLISTIC ORAL EXAM CRITERIA

Excellent: Communicates effectively and accurately with minor mistakes. Fluent and needs no assistance.

Very good: Communicates well with few mistakes. Able to use good vocabulary. More complex grammar structures. Needs almost no effort on the part of the listener.

Good: Communicates with some mistakes (vocabulary & pronunciation & grammar). Needs little effort on the part of the listener. Limited use of complex grammar structures.

Not very good: Has problems in communication. Frequent inaccuracies, little meaning conveyed.

Poor: Has serious problems in communication. Limited or no understanding and only one word simple answers.

Oral Exam Evaluation Chart

Students	Poor (0-25)	Not very good (26-45)	Good (46-70)	Very good (71-85)	Excellent (86-100)
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					

Appendix C - The Analytic Assessment Tool

	Pronunciation	Grammar	Vocabulary	Fluency	Comprehension
Very poor (0-4)	A lot of wrong pronunciation	Sentences show almost no grammatical structures	Very limited knowledge/use of English words	Hesitation and/or pause all the time	Almost no or too little understanding
Poor (5-8)	Frequent wrong pronunciation	Major problems in structures	Frequent errors in word choice	Frequent hesitation and/or pause	Limited understanding and/or needs constant repetition and rephrasing
Average (9-12)	Occasional errors in pronunciation	Several errors in constructing structures	Occasional errors in word choice	Occasional hesitation and/or pause	Understands careful and/or simplified speech with some repetition
Good (13-16)	Some errors in pronunciation	Minor problems in structure	Minor errors in word choice	Minor hesitation and/or pause	Understands normal and/or simplified speech with fewer repetition
Very good (17-20)	Minor or no errors	Few or no errors in structures	Effective/appropriate word choice	No hesitation and/or pause	Understands everything without repetition

Oral Exam Evaluation Chart

Students	Pronunciation (0-20)	Grammar (0-20)	Vocabulary (0-20)	Fluency (0-20)	Comprehension (0-20)	Total Score

Appendix D - Oral Exam Procedure Guidelines from the Oral Exam Procedure Booklet

ORAL EXAM PROCEDURE

In order to standardize the oral testing and evaluation of students, the following must be adhered to by all instructors when they give an oral exam.

The test is to be given in two phases:

Phase 1: Warming - up phase when the student is asked some basic questions about their name / surname, where they are from, etc. (approximately one minute).

Phase 2: Students are asked questions from the file provided, or other similar questions. Care is to be taken in this phase so that students are not asked about anything they have not come across in their course books (approximately 3 - 4 minutes).

- Every student, regardless of ability, must be given the same amount of time – at least 5 minutes.
- Once a student has started to give their answers, they should be encouraged to talk about the same subject, with prompts from the instructor if necessary.
- Students are not to be corrected during their oral test.
- If a student asks: "What is *(a Turkish word)* in English?", they must be given an answer. However, this is only the case if they haven't been taught that word.
- If the student is from one of your classes, the student is to be given a grade according to how they perform in the oral test, not on their classroom ability.
- Students are to be asked as many different question types as possible.
- If a student has obviously memorized 'a speech ', they should be stopped, and asked a question about something else.
- Students are not to be informed of their grades by the teacher – either at the end of the oral exam or later. All grades are posted on the SFL website.

Appendix E - Sample Questions from the *Oral Exam Procedure Booklet*

SPEAKING EXAM QUESTIONS FOR C GROUPS

1. Who do you live with?
   Have you got any brothers or sisters?
   How many?
   What do they do?
   What do they look like?
   Are you the youngest, the oldest or the middle child in the family?

2. When was your mother born?
   Where was she born?
   What does she do? If she works....
   Where does she work? Does she like her job?
   What time does she start and finish work?

3. When was your father born?
   Where was he born?
   What does he do? (What did he do? - if he is retired)
   Where does he work? Does he like his job?
   What time does he start and finish work?

4. Do you get on well with your father / mother / sister / brother?
   What kind of problems do you have with them?

5. Do you spend a lot of time with your family?
   What do you do together?

6. When is your birthday?
   Do you usually celebrate it?
   Do you remember your last birthday?
   What did you do?
   How did you celebrate it?
   Where were you?
   Who were you with?
   What presents did you get?

7. Do you spend more time with your family or with your friends?
   What is your best friend's name?
   When did you first meet him/ her?
   What is he/she like? Describe him / her.
   How often do you see him/her?
   How old is he/she?
   What does he/she do?
   What do you do together? Where did you go last together?

8. Do you live in a house or a flat?
   What is it like?
   (flat): Which floor is it on?
   (house): Does it have a garden?

9. How often do you watch TV?
   What kind of programmes do you usually watch?
   What is your favourite TV programme?
   How often do you watch it?
   What is it about?

10. Do you like going to the cinema?
    How often do you go to the cinema?
    Do you remember the last film you saw? Did you enjoy it?

11. Have you ever left the cinema in the middle of a film?

Appendix F - Transcription of a Sample Speaking Exam

Researcher: What's your name?

Student: My name's XX.

Researcher: Sefa. Sefa, where are you from?

Student: I'm from Muş.

Researcher: Ok, so you are from Muş. Which city do you like more, Kayseri or Muş?

Student: Muş

Researcher: Why?

Student: Because it is a good city and Kayseri is a bad city.

Researcher: Hmm. Ok. For example, what do you like about Muş?

Student: hmm..Because my friends are there and my family in there.

Researcher: So, for example, at the weekend, what do you do in Muş?

Student: I'm going to…şey…I go to…I go to the cinema and I go…park

Researcher: I understand. Where do you live in Kayseri? Dormitory or house?

Student: Dormitory

Researcher: Do you like the dormitory?

Student: No

Researcher: Why not?

Student: Because its meal is bad and it is careful

Researcher: Careful?

Student: Careful…Crowded

Researcher: Ha, crowded

…