

Instructors' Attitudes towards Assessing Speaking Holistically and Analytically

The Graduate School of Education
of
İhsan Doğramacı Bilkent University

by

Engin Evrim Önem

In Partial Fulfillment of the Requirements for the Degree of

Master of Arts

in

Teaching English as a Foreign Language

İhsan Doğramacı Bilkent University

Ankara

June 2015

For Dilek and İnci, my two very precious ladies...

İHSAN DOĞRAMACI BILKENT UNIVERSITY

GRADUATE SCHOOL OF EDUCATION

Thesis Title: Instructors' Attitudes towards Assessing Speaking Holistically and

Analytically

Engin Evrim Önem

Oral Defence May 2015

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Arts in Teaching English as a Foreign Language.

Prof. Dr. Kimberly Trimble (Supervisor)

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Arts in Teaching English as a Foreign Language.

Asst. Prof. Dr. Deniz Ortaçtepe (Examining Committee Member)

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Arts in Teaching English as a Foreign Language.

Assoc. Prof. Dr. Kemal Sinan Özmen (Examining Committee Member)

Approval of the Graduate School of Education

Prof. Dr. Margaret K. Sands (Director)

ABSTRACT

INSTRUCTORS' ATTITUDES TOWARDS ASSESSING SPEAKING HOLISTICALLY AND ANALYTICALLY

Engin Evrim Önem

M.A., Program of Teaching English as a Foreign Language

Supervisor: Prof. Dr. Kimberley Trimble

June 2015

The primary aim of this study is to find out language instructors' attitudes towards holistic and analytic assessment of speaking. The secondary aim is to investigate whether the scores assigned by using holistic and analytic assessment tools differed or not. Finally, this study sets out to reveal whether the scores assigned by using holistic and analytic assessment tools differed or not according to instructors' background. The research was conducted at the School of Foreign Languages, Erciyes University with twenty four language instructors between December 2014 and April 2015. An attitude questionnaire and holistic and analytic assessment tools were used to collect data.

The findings showed that the instructors in this study had different attitudes towards holistic and analytic assessment of speaking, and their advantages and disadvantages for assessment. While instructors did not have a negative attitude towards analytic assessment of speaking, they displayed a more positive attitude towards holistic assessment. Also, as a whole, the speaking exam scores assigned by holistic and analytic assessment tools did not differ. Scores were found to differ only

when instructors' years of experience was considered for scores obtained by holistic assessment. For other variables examined, there scores did not statistically differ according to the background of the instructors.

In the light of this study's findings, it can be suggested that allowing instructors to choose between holistic and analytic assessment tools of their own preference may be considered for further educational purposes.

Key words: Holistic Assessment, Analytic Assessment, English as a Foreign Language, Speaking Skill, Exam Scores.

ÖZET

OKUTMANLARIN KONUŞMA BECERİSİNİN BÜTÜNCÜL VE PARÇALI DEĞERLENDİRİLMESİNE YÖNELİK TUTUMLARI

Engin Evrim Önem

Yüksek Lisans, Yabancı Dil Olarak İngilizce Öğretimi

Tez Yöneticisi: Prof. Dr. Kimberly Trimble

Haziran 2015

Bu çalışmanın öncelikli amacı, yabancı dil okutmanlarının konuşma becerisinin bütüncül veya parçalı değerlendirilmesine yönelik tutumlarını ortaya çıkarmaktır. Çalışmanın ikincil amacı ise bütüncül ve parçalı değerlendirmenin kullanılması ile elde edilen konuşma sınav puanlarının farklılık gösterip göstermediğinin araştırılmasıdır. Son olarak, verilen puanların okutmanların art alanlarına göre değişiklik gösterip göstermediğinin incelenmesi de bu çalışmada amaçlanmıştır. Çalışma, Erciyes Üniversitesi, Yabancı Diller Yüksekokulunda 24 yabancı dil okutmanı ile 2014 Aralık ve 2015 Nisan ayları arasında gerçekleştirilmiştir. Veri toplamak için bir tutum anketi ve bütüncül ve parçalı değerlendirme araçları kullanılmıştır.

Sonuçlar, çalışmaya katılan okutmanların İngilizce konuşma becerisinin değerlendirilmesi için parçalı ve bütüncül değerlendirmeye karşı farklı tutumlara sahip olduklarını ortaya koymuştur. Bununla birlikte, çalışmaya katılan okutmanlar her ne kadar parçalı değerlendirmeye yönelik olumsuz bir tutum sergilemedilerse de, bütüncül değerlendirmeye karşı daha olumlu bir tutumları olduğu ortaya çıkmıştır.

Ayrıca, bütüncül veya parçalı değerlendirme araçları kullanılarak elde edilen puanlarda farklılık görülmemiş, okutmanların tecrübe süreleri ve bütüncül değerlendirme sonuçları dışında diğer hiçbir artalan değişkeni ile değerlendirme araçları arasında istatistiksel olarak anlamlı bir farka rastlanılmamıştır.

Elde edilen bu sonuçlar ışığında ileriki eğitimsel amaçlar için okutmanların kendi tutumları ve tercihleri doğrultusunda İngilizce konuşma becerisinin değerlendirilmesi için bütüncül veya parçalı değerlendirme araçlarından birini tercih edebilmeleri tavsiye edilebilir.

Anahtar Kelimeler: Bütüncül Değerlendirme, Parçalı Değerlendirme, Yabancı Dil olarak İngilizce, Konuşma Becerisi, Sınav Puanları

ACKNOWLEDGEMENTS

This year of MA TEFL was a real challenge and there were some very special people, without whom this challenge could not have been overcome.

First of all, I would like to express my gratitude for my thesis advisor Prof. Dr. Kimberly Trimble, who always supported me with his invaluable feedback, and guidance throughout the year and the study. I would like to thank Asst. Prof. Dr. Deniz Ortaçtepe, who made it possible for us to see this year through with her unforgettable support and friendliness. I would also like to thank to my committee member, Assoc. Prof. Dr. Kemal Sinan Özmen for his contributions to my thesis.

I am grateful to the Rector, Prof. Dr. Fahrettin Keleştemur, who gave me permission to attend this program. I am also grateful to the director of the School of Foreign Languages of Erciyes University, Fikret Kara for his support. I am indebted to my friends and colleagues at School of Foreign Languages, Erciyes University for their participation and the time they spared for me. I also would like to thank to the students who allowed me to use their speaking exams in this study.

I owe special thanks to all my classmates in the 2014-2015 MA TEFL program. We shared so much with each other. I would like to thank to İpek Dağkiran especially for being there for me when I needed with her intelligent and witty remarks.

Finally, I would like to express my appreciation and thanks to my family, who made it possible for me to survive this year. My father and mother were always with me in this long run. I owe my wife a great deal as she supported me even long before the beginning of the program. I could have never finished this program without them.

TABLE OF CONTENTS

ABSTRACT	iii
ÖZET	v
ACKNOWLEDGEMENTS	vii
TABLE OF CONTENTS	viii
LIST OF TABLES	xiii
CHAPTER I - INTRODUCTION.....	1
Introduction	1
Background of the Study.....	2
Statement of the Problem	7
Research Questions	8
Significance of the Study	8
Definition of Key Terms	9
Conclusion	10
CHAPTER II - LITERATURE REVIEW	11
Introduction	11
Assessment in Language Teaching	12
Assessing Speaking.....	14
Types of Assessment.....	16
Holistic Assessment	16
Advantages and Disadvantages of Holistic Assessment.....	16

Analytic Assessment	18
Advantages and Disadvantages of Analytic Assessment.....	18
Reliability.....	21
Raters.....	24
Factors Affecting Raters	24
Age.	25
Gender.....	25
Academic Major.....	26
Years of Experience.	27
Studies on Instructors' Attitudes towards Holistic and Analytic Assessment	28
Conclusion	29
Chapter III - METHODOLOGY	30
Introduction	30
Research Design.....	30
Setting and Participants.....	31
Instructors' Background Variables	31
Instrumentation	33
Quantitative Data Collection.....	33
Attitude Questionnaire	33
Speaking Assessment Tools.....	34
Qualitative Data Collection.....	35
Methods of Data Collection	35

Preparations for Data Collection	35
Training the Instructors	36
Data Collection Procedure	37
Methods of Data Analysis	38
Conclusion	39
CHAPTER IV - DATA ANALYSIS	40
Introduction	40
Instructors' Attitudes towards Holistic and Analytic Assessment of Speaking.....	41
Quantitative Data	41
Holistic Assessment	41
Analytic Assessment	42
Responses to Open-Ended Questions.....	44
Positive Elements of Holistic Assessment.	45
Negative Elements of Holistic Assessment.	46
Positive Elements of Analytic Assessment.	47
Negative Elements of Analytic Assessment.	48
Differences in Holistic and Analytic Speaking Exam Scores	50
Instructors' Background and the Speaking Exam Scores	50
Age and Scores.....	51
Gender and Scores.....	52
Academic Major and Scores	53
Years of Experience and Scores.....	54

Conclusion	55
CHAPTER V- CONCLUSION.....	57
Overview of the Study	57
Discussion of Findings.....	57
Research Question 1: What are the English Instructors' Attitudes towards Holistic and Analytic Assessment of Speaking?	58
Research Question 2: Is There a Difference in the Speaking Exam Scores Assigned by Instructors Using Holistic and Analytic Assessment Tools?	64
Research Question 3: Do the Speaking Exam Scores Obtained via Holistic and Analytic Assessment Differ According to the Instructors' Background?.....	65
Age and Scores.....	65
Gender and Scores.....	65
Academic Major and Scores	66
Years of Experience and Scores.....	66
Pedagogical Implications	67
Limitations of the Study.....	68
Suggestions for Further Research	69
Conclusion	70
REFERENCES.....	72
APPENDICES	77
Appendix A - The Attitude Questionnaire	77
Appendix B - Holistic Assessment Tool.....	78

Appendix C - Analytic Assessment Tool.....	79
Appendix D - Oral Exam Procedure Guidelines from the Oral Exam Procedure Booklet	80
Appendix E - Sample Questions from the <i>Oral Exam Procedure Booklet</i>	81
Appendix F - Sample Transcription of a Speaking Exam.....	82

LIST OF TABLES

Table	Page
1. The Distribution of the Instructors' Age Groups	32
2. The Distribution of the Instructors' Academic Major.....	32
3. The Distribution of the Instructors' Years of Experience.....	33
4. Results of the Holistic Assessment Questionnaire.....	41
5. Results of the Analytic Assessment Questionnaire.....	43
6. Elements Emerged from Open-Ended Items.....	44
7. Students' Speaking Exam Scores Obtained by Different Assessment Tools	50
8. Age and the Speaking Exam Scores.....	51
9. Gender and the Scores Obtained by Holistic Assessment Tool.....	52
10. Gender and the Scores Obtained by Analytic Assessment Tool	52
11. Academic Major and the Speaking Exam Scores	53
12. Years of Experience and the Speaking Exam Scores.....	54

CHAPTER I - INTRODUCTION

Introduction

Speaking is an important language skill in learning a foreign language for both students and teachers. For students, speaking is a very good sign of overall competency of 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. In fact, one of the reasons for most people to find speaking extremely difficult is this simultaneity. In this sense, it can easily be accepted that mastery in 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 the 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 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). As a result, analytic assessment tools have come to focus (Fulcher, 2003; Luoma, 2004). Analytic assessment seemed promising and useful for the precision in testing and focusing on different aspects simultaneously in

the beginning. However, it was soon understood that both analytic and holistic assessment have varying strengths and weaknesses (Brown, 2004; Luoma, 2004; Nakamura, 2004; Weir, 2005).

There are many studies about holistic and analytic assessment, each focusing on a different aspect. For example, while some studies have focused on the advantages and the disadvantages of both assessment types (Bachman & Palmer, 1996; Brown, 2004; Huot, 1990; Tuan, 2012), some focused on the reliability levels of each (Bacha, 2001; Carr, 2000; Chuang, 2009; Çetin, 2011; Nakamura, 2004; Vacc, 1989). There are some studies comparing the scores obtained by using both assessment types (Chi, 2001; Chuang, 2009; Harsch & Martin, 2012) and the effect of culture on choosing an analytic or holistic perspective (Monga & John, 2006; Nisbett & Miyamoto, 2005). However, studies about language teachers' attitudes, or in other words, considerations of actual users of both types of assessment, are limited (Chuang, 2009; Knoch, 2009). In fact, there has been no research conducted about the attitudes of language instructors towards analytic and holistic assessment of speaking in Turkey. Also, studies comparing the scores obtained by using holistic and analytic assessment of speaking are very rare to see. Therefore, this study aims to investigate the attitudes of Turkish teachers of EFL towards holistic and analytic assessment of speaking and the scores obtained by holistic and analytic assessment.

Background of the Study

Speaking is considered to be the most important skill of learning a foreign language since many people consider it as a good sign of overall proficiency. In fact, 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. With the emergence of the term “communicative competence” in the early 1970s along with the communicative approach, the emphasis put on speaking skill in English language teaching (ELT)

clearly intensified (Larsen-Freeman, 2000). However, parallel with the rise of communicative approach, how to assess speaking became a question.

Speaking is the most difficult skill to assess for many reasons. First of all, “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. In fact, Weir (2005) emphasizes the spontaneity of speaking and elaborates on the behaviors that are expected in 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)

Another controversy in assessing speaking is related to the type assessment to be used. 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 holistic assessment, the rater matches an overall impression with the descriptors to arrive at a score (p. 242). However, Fulcher (2003) points to an increasing concern about scores obtained from language tests to be meaningful. Also, the expectancy that testing scores 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 led 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 turned 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 could assess different aspects of speaking in detail, and the process taking place in the raters’ minds during assessment can be reflected to the scores easily (Bachman & Palmer, 1996; Brown, 2004). In other words, during 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).

However, although analytic assessment tools seemed promising at the beginning, debates about holistic and analytic assessment have begun to arise. For example, it was later realized that holistic and analytic types of assessment have varying strengths and weaknesses (Brown, 2001, 2004; Luoma, 2004; Nakamura, 2004; Weir, 2005). On one hand, some studies report advantages of holistic assessment compared to analytic assessment. For example, Brown (2004) highlights that holistic assessment has “relatively high inter-rater reliability” (p. 242), signaling a higher level of consistency among raters. Huot (1990) claims that holistic scoring is flexible, economical and practical and has gained acceptance widely “by employing a rater's full impression of a text without trying to reduce her judgment to a set of recognizable skills” (p. 201). Similarly, Tuan (2012) and Luoma (2004) admit that holistic scoring is more advantageous than analytic scoring when it comes to practicality since it does not require the rater to divide his/her attention into different aspects at the same time.

On the other hand, some studies suggest that analytic assessment has more advantages than holistic assessment. For instance, Brown (2004) states that analytic assessment gives a more detailed picture of the examinee since different aspects of the productive skill are analyzed. Similarly, according to Bachman and Palmer (1996), analytic scales are good at assigning levels and differentiating among weighting of components. They concluded that by using analytic assessment, every aspect of a performance is evaluated. They also argue that analytic scales provide a profile of the specific areas of language ability chosen to be tested and reflect what raters actually do when assessing the language (Bachman & Palmer, 1996). In other words, although raters seem to be paying attention to separate aspects of language in their minds during assessment, analytic assessment makes it more visible and clear to track the evaluation process (Bachman & Palmer, 1996). The debate in the literature about the advantages and disadvantages of the holistic and analytic assessment has been going on for a long time.

Another debate in the literature is about the reliability of holistic and analytic 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 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. The results in the literature are diverse in terms of the reliability issues concerning holistic and analytic assessment tools. Some research reveals that holistic and analytic assessment provided no difference in terms of scores and rater reliability. 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 reported high inter-rater reliability levels in the tests. However, some research report that using holistic and analytic assessment may affect the scores and the reliability. For instance, Carr (2000) focuses on the scores obtained via both holistic and analytic assessment and reports that changing the rating scale type has an effect on both the interpretation of that section of a test and total test scores. In another study, Nakamura (2004) reports a higher inter-rater reliability level in the analytic scoring than holistic scoring. Çetin (2011) makes a comparison between the scores obtained via analytic and holistic assessment tools and reports high inter rater reliability within the holistic and analytic scores. Some studies provide a possible solution to these varying results by offering a more integrated strategy to use holistic and analytic assessment. For example, as Harsch and Martin (2012) and Jin, Mak & Zu (2012) offer that, it is quite plausible to combine both holistic and analytic scoring to have a better assessment performance. 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.

Finally, research has also 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). For instance, Chuang (2009) reports findings that indicate statistically significant differences among speaking scores assigned by teachers with different ages and academic majors. However, no statistically significant differences are found in terms of teaching experience and scores (Chuang, 2009). In another

study, Chuang (2011) reveals that variables such as gender and especially academic background cause a certain degree of impact on test scores. In fact, Chuang (2011) states that the overall holistic scores rated by the raters with linguistics or literature major backgrounds were significantly much severe than the raters with TESOL backgrounds and other major backgrounds. In a study with contrasting results, Caban (2003) finds that rater differences do not appear to be a direct result of the raters' academic training. In other words, academic major seemed not to have a significant effect on raters' scores in that study. Consequently, results of the studies on factors such as age, gender, academic major and years of experience and their effect on the raters can be considered inconclusive.

Statement of the Problem

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 has been a shift from the traditional assessment, holistic, to analytic assessment. However, this shift also brought 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 assessment 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 (Brown, 2004; Fulcher, 2003, 2007; Huot, 1990; Luoma, 2004; Nakamura, 2004; Tuan, 2012; Weir, 2005), some focused on the scores as a result of using both tests (Bacha, 2001; Carr, 2000; Chuang, 2009; Çetin, 2011; Nakamura, 2004) and some focused on the background variables affecting scores (Caban, 2003; Chuang, 2009, 2011). However, research focusing on instructors' attitudes towards analytic and holistic grading 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.

At the School of Foreign Languages Erciyes University (EU SFL), both holistic and analytic assessment tools were used previously at different times in the past to assess speaking. For instance, after using holistic assessment for a long time, speaking exams were started to be assessed analytically. This shift brought a conflict among the instructors in terms of assessment since there were advantages and disadvantages of both types of assessment. As a result, this led instructors to have varying attitudes towards these different types of assessment and a resistance towards either one of the assessment types was seen. In this sense, EU SFL can be considered as a first hand example of the debate on holistic and analytic assessment of speaking in an ELT context. However, language instructors' attitudes towards holistic and analytic assessment of speaking and whether scores obtained by holistic and analytic assessment tools differ or not remain unclear.

Research Questions

This paper aims to find answers to the following research questions:

1. What are the English instructors' attitudes towards holistic and analytic assessment of speaking?
2. Is there a difference in the speaking exam scores assigned by English instructors using holistic and analytic assessment tools?
3. Do the speaking exam scores obtained via holistic and analytic assessment differ according to the instructors' background (age, gender, academic major and years of experience)?

Significance of the Study

This study will 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 will 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,

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 own testing processes. Also, by receiving insights from both sides of the analytic vs. holistic debate, it may also help administrators and testing professionals in other institutions to review and/or plan their testing aims.

At a more local level, the findings of this study can be used to review the assessment processes being applied at the School of Foreign Languages, Erciyes University. Since language instructors are the actual practitioners of both types of assessment, their attitudes towards using holistic and analytic types of assessment should be focused more. This way, if how teachers feel about each type of assessment can be revealed, the instructors, their institutions and learners could be satisfied with the testing performance.

Definition of Key Terms

Holistic Assessment: A type of assessment which focuses on the whole (Brown, 2004; McNamara, 2000). At the end of the holistic assessment, one overall

score is assigned either impressionistically, or guided by a rating scale for the performance (Fulcher, 2003).

Analytic Assessment: A type of assessment which is based on “assessing each aspect of a performance separately” (McNamara, 2000, p. 131). As a result of this multi-component based analysis, several scores appear at the end of the assessment procedure.

Attitude: As explained in Oxford Dictionary, attitude is a settled way of thinking or feeling about something. It is different from “perception,” which means awareness of something. Although both terms seemed to refer to the same thing at the first glance, the term “attitude” in this study is used to refer to ideas and thinking of instructors about holistic and analytic assessment of speaking.

Conclusion

In this chapter, background of the study and issues related to the assessment of speaking in teaching English were presented. Also, the research questions this study aims to answer and significance of the study as well as definition of key terms were explained in this chapter.

In the second chapter, the review of the literature and studies concerning assessment of speaking are presented. Following that chapter, methodology of the study is described in chapter three. The fourth chapter presents the procedures for data analysis and the findings of the study. The last chapter illustrates the discussion of the results and the findings, implications and limitations of the study and suggestions for further research.

CHAPTER II - LITERATURE REVIEW

Introduction

Assessing speaking is a difficult issue in language teaching because of the complexity of the nature of both the skill and the assessment process. To overcome the difficulties, different types of assessment including holistic and analytic tools have been used for a long time. However, it is later seen that both types of assessment tools have various strengths and weaknesses, which led to a debate about assessing speaking holistically and analytically. Similarly, it is quite plausible to consider that those various advantages and disadvantages of holistic and analytic assessment tools may also lead to differences among attitudes of language instructors, actual users' of those assessment tools, towards holistic and analytic assessment of speaking. However, very few research focused on language instructors' attitudes towards analytic and holistic assessment of speaking. As a result, this study aims to answer three research questions:

1. What are the English instructors' attitudes towards holistic and analytic assessment of speaking?
2. Is there a difference in the speaking exam scores assigned by English instructors using holistic and analytic assessment tools?
3. Do the speaking exam scores obtained via holistic and analytic assessment differ according to the instructors' background (age, gender, academic major and years of experience)?

This chapter presents the subject matter and related concepts in detail in respect with the literature to provide a more clear understanding of the focus of the research. After a brief introduction to language assessment, assessing speaking and

types of assessment, where holistic and analytic assessment are focused in more detail, are presented. This section is followed by the presentation of the advantages and disadvantages of holistic and analytic assessment within the light of the relevant literature, separately. A summary of studies related to the reliability issues of both holistic and analytic assessment are also presented in this chapter.

This chapter also focuses on the raters, the critical and actual users of different types of assessment. After a concise introduction, the factors affecting raters such as age, gender, academic major and years of experience are also mentioned in this chapter. This chapter ends with reporting on limited studies about instructors' attitudes towards holistic and analytic assessment.

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 as “a universal feature of social life” (McNamara, 2000, p. 3). People are tested about different topics every day, such as driving tests or achievement tests, to show that they meet some 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, assessment in language is the process of measuring an individual’s skills and/or competence in language. As a result, tests can be considered as valuable tools and a subset for assessment, which is an indispensable part of language teaching (Brown, 2004).

There are several reasons why tests and assessment 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 clearly be seen that tests are important elements of language teaching and learning.

Since testing is an important part of education, the twentieth century saw some major changes in approaches towards teaching and assessment in language education, which still prevail today (Brown, 2004). The first approach towards teaching and assessment was “the separate units of language approach”. According to Bachman and Cohen (1999), the dominant view of language ability in 1960s and 1970s was derived from a structuralist linguistic view, which accepted 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 component parts and that those parts can be tested successfully” (p. 8).

In response to the structural view, the integrative approach was an alternative and the second major approach seen towards teaching and testing in the 20th century. This approach suggested an indivisible view of language proficiency (Brown, 2004) with the 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 were expected to integrate different pieces of linguistic knowledge during assessment (Brown, 2004).

As the language teaching and assessment field developed in the 1980s, another approach towards assessment, a competence model based on transformational-generative linguistics, that combined the user’s knowledge with performance appeared (Bachman & Cohen, 1999; Brown, 2004). This new paradigm resulted 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, led to a quest for authenticity and this was 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 became as important as receptive skills like reading and listening which were implied in the first two approaches.

Assessing Speaking

Assessing productive skills, especially speaking, is more difficult than receptive skills for several reasons. First, the basic nature of speaking, interaction, is 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 assessment of speaking.

A second reason for the difficulty is related to the spontaneity of speaking. Speaking is a spontaneous production skill and therefore, during assessment, “participants have to produce their own language according to their own resources” (Erlam, 2009, p. 65). In such circumstances, the exam context becomes primary and meaning along with time pressure affect the context and the performance (Erlam, 2009). This issue raises two questions. First, there are many studies revealing 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 to an inaccurate evaluation of speaking. Secondly, the 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 scoring procedure as well as the raters are among the other difficulties seen during assessment of speaking. As for the scoring procedure, the type of assessment and the tool to be used accordingly are equally difficult to choose since different types of assessment have varying strengths and weaknesses. Also, raters are the key figures in assessment. 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 assessment. In the further sections, these issues are dealt with more details.

Types of Assessment

Although there are different types of assessment, this study focuses on only holistic and analytic types of assessment of speaking.

Holistic Assessment

In simple terms, holistic assessment, also called as traditional assessment (Lumley, 2005), focuses on the whole (Brown, 2004; McNamara, 2000). 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.

Advantages and disadvantages of holistic assessment. Holistic assessment has both advantages and disadvantages. Within the literature, several advantages of holistic assessment are discussed. According to Brown (2004) and Weir (2005), speed is one of them. Since raters do not need to focus on separate components of a performance during 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 reflect their own 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 holistic assessment. It is plausible to think that without being limited to a narrow set of criteria, holistic assessment 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)

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). 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 towards 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.

Analytic Assessment

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, Clapham and Wall (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, Clapham & Wall, 1995), which depends on the requirements and/or expectations of the institution.

Advantages and disadvantages of analytic assessment. Analytic assessment also 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, scorers 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 scorers. 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, analytic assessment makes the assessment process clear to the 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, differentiating and weighing components in accordance with the expectations (Alderson, Clapham & Wall, 1995; Bachman & Palmer, 1996). This *differentiation* and *weighing* refers to the flexibility to change each components' score weight or 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 the 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, analytic assessment may be more useful for diagnostic purposes (Luoma, 2004; Weir, 2005). As analytic assessments focus on multiple components of a performance, it 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, 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 for analytic assessment is that it limits the freedom of the raters. As Fulcher (2010) mentions, teachers tend to have negative attitudes towards highly detailed test specifications because such specifications are limiting. As teachers are required to follow certain and specific guidelines during 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 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 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 positive or negative effect on the proceeding score of the component. In addition to positive or negative effect, 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, 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 when assessment is considered. 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 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, there are studies reporting that changing the rating scale type, namely, from holistic to analytic or vice versa has an effect on scores. For example, as Carr (2000) reports on the findings in her study, 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.

The whole reliability issue related to the type of assessment is controversial in the literature. On the one hand, it is possible to find studies reporting a change in the reliability level depending on the assessment tool being used. For instance, 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 suggest a higher reliability level in analytic assessment. Similarly, Nakamura (2004) reports a higher inter-rater reliability level for the analytic scoring than holistic scoring. In contrast, some studies report a higher level of reliability in favor of holistic assessment. For instance, Barkaoui (2007) compares the holistic and analytic scores of 24 EFL essays and reports that holistic scoring shows a 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 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.

On the other hand, 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 statistically 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. He 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, more research is required.

Raters

Within the literature the term, *raters*, is frequently used interchangeably with *assessors* or *scorers*. Yet, *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 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 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 seem to work to some extent, total elimination of differences in the scores seem impossible (Wang, 2010; Weigle, 1994) and “not all human elements can be compromised in the evaluation process” (Vanniarajan, 2006, p. 290). In fact, as a solution for this problem, Wang (2010) suggests creating *rater files* that include information for each and 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 study 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. Age is closely related to the years of experience in teaching. However, studies comparing scores of teachers with different age groups in ELT are limited. For example, Chuang (2009) compared scores assigned to the speaking performance of EFL learners by four groups of teachers with different ages. The results showed a statistically significant difference among groups with the youngest raters in the study (21-30) scoring the lowest while the oldest group (50+) scored the highest. In other words, the raters of the age group over 50 were the most lenient in terms of scoring and the younger teachers become the more severe scores they seem to assign.

Gender. The literature on the effect of the raters’ gender 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) analyzed the scores of male and female raters on oral performance of EFL learners. The results showed a statistically significant difference between the scores, with male raters scoring higher than female raters. Similarly, Porter (1991) found that among a range of variables, including the personality of the participants and their degree of acquaintanceship, the only variable that had 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 those of male raters. For example, in a study by Gholami, Sadeghi and Nozad (2011), both male and female raters were 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 were high, students receive higher scores from the female raters and the difference was statistically significant (Gholami, Sadeghi & Nozad, 2011). Similarly, O'Sullivan and Porter (1996) focused on the interview scores of ESL learners assigned by male and female raters and reported that female raters' scores were statistically significantly higher in that case.

Yet, some studies reported findings which 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 was seen that gender does not have a significant effect on the scores. Similarly, Chuang (2011) reported no statistically significant difference in terms of the scores assigned by the male and female raters. In other words, gender of the raters did 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 higher score. For example, Buckingham (as cited in Chuang, 2011) revealed that male students got higher scores when being interviewed by a male rater or vice versa. To sum up, regardless of all different results in the literature, the gender of the rater seems to be a variable that may affect the scores assigned.

Academic major. Academic major or the department graduated from is another background variable to be investigated in this study. In a study, Chuang (2009) analyzed the scores of raters with different academic majors (literature/linguistics, TESOL/ESL/ EFL, others). Results showed that a statistically significant difference was seen among the scores of the raters whose academic major is literature/linguistics and TESOL/ ESL/EFL. Scores of the raters with literature/linguistics major were the lowest while the scores of the raters with teaching major were the highest and the results were consistent within the groups.

However, the difference between these two groups and linguistics/literature was not significant (Chuang, 2009). This result may suggest that because the expectations of the literature/linguistics major raters were high, they tended to score low. On the other hand, raters with teaching majors may be considered as more tolerant to errors than others and this tendency was reflected in their scores. In another study, Chuang (2011) asked teachers with different majors such as literature, linguistics and teaching to evaluate the speaking performances of 75 students EFL students using a holistic assessment tool. The results revealed that academic background had a certain degree of impact on the test scores assigned by the raters. Deeper analysis showed that the scores of the raters with linguistics or literature majors were significantly lower than the raters with TESOL backgrounds and other major backgrounds.

Years of experience. Years of experience is an important 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) conducted a study about English teachers' scoring performance for speaking and focused on the effect of teaching experience and other background. The results indicated no statistically significant difference between years of experience, which meant the scores of the teachers for speaking assessment did not differ in terms of years of experience in that EFL case. However, in another study by Song and Caruso (1996), it was found that raters with more years of experience had a tendency to be more lenient when using holistic assessment tools. In a more recent study, similar results were seen. Huang and Jun (2015) focused on raters determining native-likeness of the speech samples and compared three groups of raters' (inexperienced, experienced and advanced) scoring performance for speech production. The findings showed that inexperienced raters were 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 have an effect on the scores they assign.

Studies on Instructors' Attitudes towards Holistic and Analytic Assessment

Although language instructors are the actual user of assessment tools, there are only a few studies in the literature on the attitudes of the instructors towards holistic and analytic assessment. Of the limited studies, Knoch (2009) focused on teachers' perceptions towards a scoring scale with less specific descriptors and a scale with more detailed descriptors. The results showed that raters preferred to use the more detailed, analytic scale. The main reason for that was the raters' in that study believed that the more detailed scale helped them to focus more on the details rather than an overall impression. Interestingly, most of the raters considered the more detailed scale as "minimally more time consuming" (Knoch, 2009, p. 298) and some even considered it faster.

In another study, Barkaoui (2010) compared novice and experienced teachers' and their perceptions of holistic and analytic scales. He found that novice and experienced teachers had different perceptions of the use for the scales. Barkaoui (2010) summarized 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 found that novice teachers had a tendency to refer to the rating scales more frequently than the experienced raters. In other words, novice teachers checked themselves whether they are consistent with the scales more often than the experienced teachers do. As Barkaoui (2010) concluded, “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 minds. Similarly, less experienced teachers may lack such criteria and may be dependent on outer sources during assessment. Yet, as limited studies suggest, there is a difference in terms of the teachers’ perceptions towards using the holistic and analytic assessment tools.

Conclusion

Speaking is a difficult skill to assess because there are many factors affecting such as the rating type and the raters. In this chapter, the literature related to the assessment of the speaking was presented. After presenting the relationship among speaking and other skills, types of assessment, specifically holistic and analytic assessment, were introduced. The advantages and disadvantages of holistic and analytic assessment were given and following that, raters and the factors affecting raters were focused in this chapter. The next chapter introduces the methodology of the study.

Chapter III - METHODOLOGY

Introduction

As there are different types of assessment of speaking such as holistic and analytic with different advantages and disadvantages, it is quite plausible to consider that language instructors have varying attitudes towards each type of assessment. Therefore, this study investigates English instructors' attitudes towards holistic and analytical assessment of speaking. An additional aim of this study is to explore if the speaking assessment scores differ according to the assessment tool (holistic or analytic) used. Finally, if the scores assigned by using holistic and analytic assessment tools differ according to the background variables of the instructors (age, gender, academic major and years of experience) are also focused in this study.

This chapter presents the methodology used to answer the research questions. 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 includes both quantitative and qualitative data collection methods, and methods of data analysis are presented.

Research Design

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 which 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 were 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 Erciyes University, School of Foreign Languages (EU SFL) were used to obtain quantitative data about the attitudes of the instructors towards speaking exam assessment. Two open-ended items which were included at the end of the attitude questionnaire to gain deeper insights into the attitudes of the instructors were the main source of qualitative data in this study.

Setting and Participants

The study took place between December 2014 and April 2015 at Erciyes University, School of Foreign Languages (EU SFL), Kayseri, Turkey over a period of five months. Since both holistic and analytic assessment of speaking were used previously at SFL, instructors who experienced using both assessment types of speaking were included only as participants. As a result, twenty four full time language instructors working at SFL participated in the study voluntarily, although there are currently over 100 language instructors at the EU SFL.

Instructors' Background Variables

The background variables of the instructors who participated in this study included 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 were male and ten of them were female. As for the ages of the instructors, their ages ranged from 26 to 46+. The highest number of participants for the age groups was in the 31-35 years old group. The frequency for the age groups of the instructors is given in Table 1.

Table 1

The Distribution of the Instructors' Age Groups

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

Table 2

The Distribution of the Instructors' Academic Major

Academic Major	N
ELT	13
Literature	9
Linguistics	2

The instructors had varying years of experience in teaching English and they were grouped in five categories (since all the instructors who participated in the study had more than at least five years of experience). Instructors with 11-15 years of experience composed the largest group in terms of years of experience. The distribution of the years of experience can be seen in Table 3.

Table 3

The Distribution of the Instructors' Years of Experience

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

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

Instrumentation

Both quantitative and qualitative data collection tools were used in the study to obtain data related to the attitudes of the instructors along with speaking exam scores.

Quantitative Data Collection

Attitude Questionnaire

After reviewing the literature about analytic and holistic assessment (Bachman & Palmer, 1996; Brown, 2004; Fulcher, 2003; Hughes, 2003; Huot, 1990; Luoma, 2004; Weigle 2002; Weir, 2005), the researcher created a questionnaire depending on the advantages and disadvantages of both types of assessment. The four-point Likert questionnaire consisted of 20 multiple choice items which were designed to reveal the instructors' attitudes towards the holistic and analytic assessment of speaking concerning the advantages such as ease of use or effectiveness of each type of assessment. The questionnaire was 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) were used to reveal the attitudes of the instructors towards holistic assessment of speaking and items # 3 and

20 were reverse coded. As for the items related to analytic assessment, items # 1, # 4, # 6, # 7, # 10, # 11, # 12, # 14, # 17 and # 18 were used and items # 12 and # 14 were 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 were re-coded in SPSS.

The data were entered into SPSS 22 for the two sets of questions and a test of internal consistency was run to see whether the questionnaire had high internal consistency. The *Cronbach's alpha* for each set of questions was calculated as .93 for holistic assessment and .92 for analytic assessment. As a result of consulting to an expert and high numbers of internal consistency, the questionnaire was accepted as reliable (see Appendix A). A short section of demographic data concerning the instructors' background (gender, age, years of experience and academic major) was also included in the beginning.

Speaking Assessment Tools

For assessing speaking performances of students, holistic and analytic tools assessment tools were used in the study. Both tools were developed and validated by the *Testing Office* of EU SFL, which is responsible for test administrations and evaluations of the scores. After designing and using the holistic assessment tool for nearly ten years, the *Testing Office* of EU SFL created the analytic assessment tool based upon the relevant literature and the needs of the EU SFL. Since both tools were used previously for assessing speaking exams given at EU SFL, the tools were considered as reliable and valid for this study as well.

The holistic assessment tool included a basic set of descriptors for raters to follow during assessment and a range of scores for performance of the student, among which the raters were asked to assign only one score they feel appropriate for the overall performance (see Appendix B). The analytic assessment tool, on the other

hand, included different components of speaking such as pronunciation, grammar, vocabulary, fluency and comprehension for raters to assess the learners' speaking performance. It had a more detailed set of descriptors for each component of speaking separately. It also included a range of scores for each component of speaking for the raters to choose to assign and the final score was calculated by adding up the scores assigned for each component (see Appendix C).

Qualitative Data Collection

Along with the multiple choice items included in the attitude questionnaire, two open-ended items designed for the instructors to write the reasons about their positive or negative attitudes for both types of assessment were also included at the end of the questionnaire. The items were semi-structured since they required 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).

Methods of Data Collection

Preparations for Data Collection

The researcher obtained the holistic and analytic assessment tools which were used previously at EU SF prior to the data collection process as well as the *Oral Exam Procedure Booklet* from the *Testing Office* of EU SFL (see Appendix D for the guidelines), which included directions to be paid attention during speaking exams. These directions included 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 EU SFL volunteered to participate in the study. Since each student was placed in the pre-intermediate classes after the proficiency exam given previously at EU SFL at the beginning of the Fall term of 2014/2015, they were all regarded as students with similar levels of English. These students were

given a speaking exam individually by the researcher in December 2014 in two days and the exams took place at the office of the researcher at EU SFL. The speaking exams took approximately five minutes and they consisted of a warm-up section (approx. one minute) and a question section (approx. four minutes). All the questions asked during the speaking exam were chosen from the *Oral Exam Procedure Booklet* (see Appendix E for sample questions). The exams started as normal conversations (warm-up section) and then the researcher asked questions from the *Oral Exam Procedure Booklet*. Since the level of the students was pre-intermediate, the researcher assisted the students when they needed help by rephrasing the questions, changing the topic or asking follow up questions. Although there was not a constant number for the questions asked, students were asked approximately seven to ten questions. During the speaking assessment, all of the questions and students' oral responses were recorded using a sound recorder by the researcher as an audio file (see Appendix F for a transcribed sample of the exam).

Training the Instructors

Due to the large number of students and feasibility issues, EU SFL stopped assessing speaking performances of the learners five years ago. Therefore, although only the instructors who were familiar with using both assessment tools participated in this study, they were nevertheless trained to use both analytic and holistic assessment tools separately by the researcher. Each training session took place in the conference hall of the EU SFL and each session lasted for an hour. The instructors were trained to use the holistic assessment tool in January first. During the training session, the researcher explained the rubric to the instructors and gave information about the assessment process. In April, the researcher held another training session for the analytic assessment tool. Similar to the training session for the holistic assessment, the rubric was explained and information about the assessment process

was given. All of the instructors participated in the assessment process attended both training sessions.

Data Collection Procedure

Ten randomly selected prep class students studying English at Erciyes University, School of Foreign Languages were recorded separately during a 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 had experience in using both types of assessment tools were trained to use the holistic assessment tool in January. At the end of the training session, all of the instructors who participated in the study listened 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 of EU SFL. While listening to the recordings, they were 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 took approximately fifty minutes and the tools and the scores given were collected at the end of the process.

Later, a similar procedure was followed for the analytic assessment of the speaking exams. However, in order to minimize the instructors' recalling of the speaking exams and the scores they assigned, the analytic assessment of the speaking exams was held a month later. As a result, at the end of the training session in April, while all of the instructors who participated in the study were present at the conference hall of EU SFL, analytic assessment of the speaking exams took place. The instructors listened to the audio recordings of the same speaking exam together at the same time only once without any interruptions again and used the analytic assessment tool to assess the speaking performances of the students provided by the

researcher. The process took approximately fifty minutes and at the end of the process, the tools and the scores given were collected again.

Following the week of the analytic assessment, the instructors were administered the attitude questionnaire prepared by the researcher, which was composed of 20 four-point Likert Scale items and two open-ended questions, to reveal the instructors' attitudes towards holistic and analytic assessment. They were asked to choose the option they feel reflect their attitudes towards holistic and analytic assessment at best and provide the reasons for their choices as much detailed as possible through the open-ended items. The instructors filled in the questionnaires in their own spare time and the completion of the questionnaire took a week.

At the end of the data collection procedure, both quantitative and qualitative data were obtained. The quantitative data consisted of the scores obtained from the attitude questionnaire and the speaking exam scores obtained from holistic and analytic assessment tools. The qualitative data included the answers to the open-ended questions for the instructors' reasons for their attitudes towards different types of assessment.

Methods of Data Analysis

To reveal the attitudes of the instructors towards holistic and analytic assessment of speaking, the quantitative data obtained from the attitude questionnaire were analyzed by descriptive statistics while the qualitative data were analyzed thematically. Means and standard deviations for each item was calculated and presented. The thematic analysis was based on the frequency of the positive and negative elements that were seen in the responses of the participants and similar elements were grouped under same themes. To see whether the speaking exam scores assigned by using different types of assessment differed or not, a paired samples *t*-test was 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 was used.

Conclusion

In this chapter, the methodology of the study was presented. After explaining the research design, the setting and the participants of the study were introduced. Following that, the instruments used in the study as well as the data collection methods and methods of data analysis were also provided in detail. In the next chapter, the analysis and results of the data will be presented.

CHAPTER IV - DATA ANALYSIS

Introduction

The primary purpose of this study was to find the attitudes of language instructors towards holistic and analytic assessment of speaking skill. Secondly, the study examined whether the scores assigned for speaking using holistic and analytic assessment tools differed or not. Additionally, this study also focused on determining whether the scores assigned for speaking via holistic and analytic assessment tools differed or not according to the instructors' background. Therefore, this research was designed to find the answers to the following research questions:

1. What are the English instructors' attitudes towards holistic and analytic assessment of speaking?
2. Is there a difference in the speaking exam scores assigned by English instructors using holistic and analytic assessment tools?
3. Do the speaking exam scores obtained via holistic and analytic assessment differ according to the instructors' background (age, gender, academic major and years of experience)?

The results of the data analysis are presented in three separate sections in this chapter. In the first section, the results of the analyses of the questionnaire regarding the attitudes of the instructors towards holistic and analytic assessment are presented. Also, the qualitative data analyses related to the attitudes of the instructors are explained in the same section. In the second section, the analysis of the speaking exam scores obtained via holistic and analytic assessment tools is explained. Finally, in the third section, the results of the analyses concerning the speaking scores and the background variables are given.

Instructors' Attitudes towards Holistic and Analytic Assessment of Speaking

In this section, first, the instructors' attitudes towards holistic and analytic assessment of speaking are presented separately regarding the four point Likert Scale items of the questionnaire. Then, the analysis of the qualitative data obtained via the open-ended items at the end of the questionnaire is explained at the end of this section.

Quantitative Data

Holistic Assessment

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

Table 4

Results of the Holistic Assessment Questionnaire

Item #*	$\bar{x}$	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 is assigned to *Strongly Agree*, 3 is assigned to *Agree*, 2 is assigned to *Disagree* while the lowest score (1) is assigned to *Strongly Disagree*. Reverse coded items were re-coded in SPSS.

The descriptive analysis of the results of the attitudes for holistic assessment showed that item # 3, *Holistic assessment is time consuming*, received the highest mean score ($\bar{x} = 3.37$, $SD = .711$). It should be kept in mind that the item was a reverse coded item and it was re-coded. In this sense, the participants had the most positive attitude towards this item, implying that holistic assessment was not time consuming. The second most positive attitude for holistic assessment was seen in two items: # 13, *Grading with holistic assessment is easier than using analytic assessment* ($\bar{x} = 3.21$, $SD = .658$) and # 5, *Holistic assessment enables rater to score more freely* ($\bar{x} = 3.21$, $SD = .779$). Following these items, item # 19, *Holistic assessment is better than analytic assessment in terms of practicality* ($\bar{x} = 3.08$, $SD = .776$), item # 9, *Speaking performance should be assessed as holistically* ($\bar{x} = 3.00$, $SD = .978$) and item # 2, *With holistic assessment, I feel the scores I assign reflect my true impressions* ($\bar{x} = 3.00$, $SD = .780$) received positive scores from the instructors. Of the final four items, which received lower scores than the previous items, there was a slight difference in terms of their means. For instance, the attitudes of the instructors were slightly lower for # 15, *I feel more confident and comfortable when assessing performance with a holistic assessment tool* ($\bar{x} = 2.96$, $SD = .806$), # 20, *At the end of the holistic assessment, the final score is higher or lower than I expected* ($\bar{x} = 2.92$, $SD = .776$) and # 8, *Holistic assessment allows the rater to reflect his/her impressions of the performance better* ($\bar{x} = 2.88$, $SD = .850$). Item # 16, *Better feedback for students can be provided with holistic assessment* ($\bar{x} = 2.42$, $SD = .881$) received the least positive score.

Analytic Assessment

The questions related to the analytic assessment consisted of ten items as well. To reveal instructors' attitudes towards analytic assessment for speaking, items

1, # 4, # 6, # 7, # 10, # 11, # 12, # 14, # 17 and # 18 were used. Items # 12 and # 14 were reverse coded. The results can be seen in Table 5.

Table 5

Results of the Analytic Assessment Questionnaire

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

performance with more freedom ($\bar{x}$ = 1.83, SD = .702) and # 12, *Analytic assessment takes a long time* ($\bar{x}$ = 1.79, SD = .779).

Responses to Open-Ended Questions

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

Table 6

Elements Emerged from the Open-Ended Items

	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

Table 6 (continued)

Elements Emerged from the Open-Ended Items

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 for separate aspects
---------------------------------	--	--

The elements are discussed in more detail separately in the following sections.

Positive elements of holistic assessment. Several themes were observed in the instructors' responses to the open-ended questions. The most common theme to emerge in the attitudes of the instructors was *the practicality of holistic assessment*. Eleven out of fourteen of the instructors who had positive attitudes towards holistic mentioned that holistic assessment was more practical. One of the instructors suggested that it was because of its convenience for their (assessment) purposes. The instructor went on to explain that due to the high number of students to take the exams in a limited time, it was easier to assign only one overall score. One other instructor highlighted that it was difficult to follow many criteria and the student at the same time and that's why holistic assessment was more practical. As a result, it can be inferred that the instructors with positive attitudes towards holistic assessment believed that the practicality of the assessment tool was an important issue.

Another similar theme seen was *the idea for assessing speaking performance holistically*. One of the instructors supported 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 was bigger

than the pieces. In a similar vein, another instructor claimed that holistic assessment was more realistic and practical than analytic assessment because people took assessments as a whole in real life and “so should speaking be”.

Another theme observed in the open-ended items was the ability to give *true reflections of the raters' impressions with holistic scoring*. One of the instructors wrote that s/he felt 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 emphasized the freedom holistic assessment creates and wrote “I feel that I have given the most accurate score by using holistic scoring since you are more free”.

Negative elements of holistic assessment. Several different themes emerged in the instructors' criticisms towards holistic assessment. The first theme seen was *the subjectivity of holistic assessment*. As one of the instructors wrote, “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 shared similar ideas went 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, she mentioned that she did not feel that the scores obtained via holistic assessment reflected her true impressions. These comments suggested that clearer guidelines were needed to ensure the reliability of the scores.

Another theme observed had a similar concern about *the vagueness the rating process*. One of the instructors criticized the holistic assessment as a very vague type of assessment, by which no clear and visible assessment process could be achieved. Another instructor wrote “it (holistic scoring) does not provide concrete reasons for

the grade I assign to my students” and emphasized the issue of subjectivity and vagueness of holistic assessment process one more time.

One final critical theme of holistic assessment was *the necessity of training instructors to use holistic assessment*. One of the instructors wrote that unless teachers were trained about how to use the holistic criteria, the scores could not be considered as reliable. Another instructor also pointed to rater training and stated 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 were critical of holistic assessment for being vague and subjective.

Positive elements of analytic assessment. Just like the themes that emerged for holistic assessment, different but related themes were also seen in the themes for analytic assessment. The most common positive element of analytic assessment was *the feedback opportunities* that it provided. Seven out of ten instructors with a positive attitude towards analytic assessment agreed that analytic assessment provided feedback opportunities for the students. Some of the instructors mentioned that this could help students to see the specific speaking performance points they needed to improve using analytic assessment.

The second theme observed was *the reliability of the scores* obtained by analytic assessment. For instance, instructors seemed to agree that analytic assessment tools produce scores with a higher reliability. Although there was not any specific information provided by the instructors, the term “reliability” was assumed to be referring to both inter and intra reliability of the raters. For instance, one of the instructors stated that since “you have a more detailed series of guidelines, the reliability of the scores assigned would be higher”. Another instructor pointed to the need for guidelines to reach more reliable assessment results. As a result, the

instructors with positive attitudes towards analytic assessment seemed to pay attention to the reliability issues in assessment.

The final positive element in the instructors' responses towards analytic assessment was *the comfort of analytic assessment*. The instructors who explained their reasons for their positive attitude in the open ended items stated that they were 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 had a positive attitude towards analytic assessment seemed to believe that having clear cut guidelines was very useful. In sum, those who viewed analytic assessment positively seemed to do so because of the feedback it provided to students, the reliability of the scores, and their comfort with specific scoring guidelines.

Negative elements of analytic assessment. The most common criticism that emerged in the instructors' responses was the quantity of *the time needed to complete analytic assessment*. Nearly all of the instructors who indicated negative attitudes towards analytic assessment emphasized this theme clearly. For instance, one of the instructors wrote "it (analytic assessment) takes a lot more time than it should" and another one wrote "focusing on different aspects takes a lot of time and is not practical". Another instructor wrote that analytic assessment was very time consuming because it needed to focus on different aspects at the same time and as a tiring type of assessment.

The second theme that emerged was the emphasis of *assessing speaking as a whole, not piece by piece*. For example, one of the instructors wrote "the quality of pieces does not make any sense for me if the whole body does not function/work", signaling his/her positive attitude for holistic assessment. Another instructor warned

that during 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 was *the inability of reflecting the true impressions of the rater because using analytic assessment*. In fact, one of the instructors mentioned that “analytic assessment makes scoring complicated and this prevents the rater to reflect his/her impression of true performance”. A second instructor explained that analytic assessment was misleading in terms of the final score. It can be assumed that the instructors who suggested these ideas were referring to adding up the scores of separate sections to reach a final score at the end of analytic assessment. They seemed to believe that as a result, the final score might be higher or lower than what had been anticipated.

Similarly, *dividing attention for separate aspects* was the final theme emerged. One of the instructors wrote “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 wrote “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 suggested 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 seemed to revolve around time consumption, the difficulty in reflecting the raters’ real impressions of the performance and the difficulty in assessing different parts simultaneously.

Differences in Holistic and Analytic Speaking Exam Scores

To answer the second research question, whether speaking exam scores assigned by holistic assessment differ from scores using analytic assessment tool, means of the speaking exam scores obtained via holistic and analytic assessment tools were 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

Students' Speaking Exam Scores Obtained by Different Assessment Tools

Student #	Holistic Score		Analytic Score	
	$\bar{x}$	SD	$\bar{x}$	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 did not yield a statistically significant difference between the scores obtained by holistic assessment tool ($\bar{x}$ = 56.80, SD = 8.44) and analytic assessment tool ($\bar{x}$ = 57.26, SD = 9.76, $t(23) = .190$, $p > .05$) and there was 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 was used to assess speaking performance at the sample speaking exams in this study, the scores assigned were nearly the same.

Instructors' Background and the Speaking 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 instructors' background, the means of the scores assigned for the speaking exams via

holistic and analytic assessment tools were compared in accordance with the age, gender, academic major and years of experience of the instructors. The results of the analyses are presented separately.

Age and Scores

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

Table 8

Age and the Speaking Exam Scores

		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 it can be seen in the table, results showed no statistically significant difference among different age groups for the scores they assigned 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 assigned similar scores for the sample speaking exams when using both assessment tools.

Gender and Scores

The data obtained from the holistic and analytic tools for speaking exams were compared according to gender by employing an independent samples *t*-test. The results of *t*-test regarding gender and the speaking exam scores obtained by holistic assessment tool are presented in Table 9.

Table 9

Gender and the Scores Obtained by Holistic Assessment Tool

Holistic Score	Male		Female		N	95% CI for Mean Difference	r	t	df
	$\bar{x}$	SD	$\bar{x}$	SD					
	54.26	6.31	60.36	10.01	24	-13.00, .808	.36*	1.83*	22

* $p > .05$

The results showed that for the scores assigned by using holistic assessment, there was not a statistically significant difference between male participants ($\bar{x} = 54.26$, $SD = 6.31$) 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 assigned similar scores using holistic the assessment tool.

The results of the *t*-test for gender and the speaking exam scores obtained by analytic assessment tool can be seen in Table 10.

Table 10

Gender and the Scores Obtained by Analytic Assessment Tool

Analytic Score	Male		Female		N	95% CI for Mean Difference	r	t	df
	$\bar{x}$	SD	$\bar{x}$	SD					
	58.35	9.06	55.74	10.97	24	-5.88, 11.10	.13*	.638*	22

* $p > .05$

Similarly, the results showed that for the scores obtained by using analytic assessment tool, there was 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 assigned similar scores using analytic assessment tool as well.

Academic Major and Scores

Participants' academic majors varied from ELT, Literature and to Linguistics departments. The data obtained from the holistic and analytic assessment tools for speaking exam scores were 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

Academic Major and the Speaking Exam Scores

		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 showed that there was not a statistically significant difference among participants' academic majors and the scores they assigned 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 assigned similar scores using holistic assessment and analytic assessment in the present study.

Years of Experience and Scores

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

Table 72

Years of Experience and the Speaking Exam Scores

		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 showed that there was a statistically significant difference among participants' years of experience and the scores they assigned using holistic assessment tool [$F(3, 20) = 3.905, p < .05$] and the effect size was large $\eta^2 = .37$ between groups. The result of a post hoc LSD test indicated that only the differences between 6-10 and 11-15 and 16-20 and 20+ were not statistically significant. However, when the groups with different years of experience were analyzed closer, the less experienced instructors seemed to score higher. For example, instructors with 6-10 years of experience had the highest mean scores ($\bar{x} = 62.05, SD = 7.79$) and was followed by the 11-15 year experience group ($\bar{x} = 61.13, SD = 8.24$). However, the scores assigned by the other two groups of 16-20 ($\bar{x} = 51.28, SD = 4.22$) and 20+ ($\bar{x} = 51.42, SD = 6.89$) were the lowest. On the other hand, there was not a

statistically significant difference in the scores obtained via analytic assessment tool and the years of experience of the instructors [$F(3, 20) = 1.998, p > .05$].

Conclusion

The analyses of the data obtained through the attitude questionnaire and holistic and analytic assessment tools were presented in this chapter. In the first section, it was seen that instructors had varying attitudes for holistic and analytic assessment and the mean scores obtained from the attitude questionnaire were higher for the holistic assessment than analytic assessment. In fact, for the items related to the holistic assessment, *strongly agree* or *agree* were the most common choices the instructors in the study chose, while only *agree* and *disagree* were common among the responses of the instructors for analytic assessment. In depth analyses of the qualitative data showed that the most common positive elements emerged for the holistic assessment included *practicality*, *necessity for assessing a performance holistically* and *true reflections of the raters via holistic scoring*. However, *the subjectivity*, *the vagueness of the rating process* and *the necessity of training instructors to use holistic assessment* were the themes emerged for criticizing holistic assessment. On the other hand, themes that emerged related to analytic assessment were *the rich feedback opportunities it provided*, *the reliability of the scores* and *its comfort*. However, *the amount of time for scoring*, *the necessity to assess speaking as a whole*, *the gap between the perceived impressions of the rater and the calculated score* and *the cognitive demands to assess separate aspects of the performance* emerged as a criticism of analytic assessment. As for the scores assigned via either assessment tool, there were not any statistically significant differences among the scores. Similarly, except for the only statistically significant difference in terms of years of experience for the scores obtained by holistic assessment tool, there were no statistically significant differences between the scores obtained via holistic or

analytic assessment tool in terms of the background variables of the participants. The next chapter will focus on the further discussion of the findings in the light of the relevant literature.

CHAPTER V- CONCLUSION

Overview of the Study

The primary aim of this study was to find the attitudes of language instructors towards holistic and analytic assessment of speaking skill. The secondary aim was to find whether the scores assigned for speaking via holistic and analytic assessment tools differed or not and whether the scores differed according to the background of the instructors.

This study addressed the following three research questions:

1. What are the English instructors' attitudes towards holistic and analytic assessment of speaking?
2. Is there a difference in the speaking exam scores assigned by English instructors using holistic and analytic assessment tools?
3. Do the speaking exam scores obtained via holistic and analytic assessment differ according to the instructors' background (age, gender, academic major and years of experience)?

In this chapter, the major findings of the study for the research questions are summarized and presented. Pedagogical implications of the findings and the limitations of the study are also discussed in this chapter. Finally, suggestions for further research are included at the end of this chapter.

Discussion of Findings

The findings of the study are presented in this chapter in three separately for each research questions. In the first section, the findings related to the instructors' attitudes towards 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.

Research Question 1: What are the English Instructors' Attitudes towards Holistic and Analytic Assessment of Speaking?

The analysis of the results for both the four-point and semi-structured open ended items in the attitude questionnaire showed that instructors in this study had different attitudes towards holistic and analytic assessment.

First, except for one item, the mean scores for all the items of holistic assessment were higher than the mean scores for the nine out of ten items of analytic assessment. In other words, the results of the questionnaire showed that the instructors who participated in the study showed a more positive attitude towards holistic assessment of speaking.

Secondly, when the items in the questionnaire and the positive and negative themes emerged from the responses to the open-ended items were analyzed, several overarching themes were seen, mostly in favor of holistic assessment. The most common support for holistic assessment and criticism for analytic assessment seen was the issue of *true reflections of the impressions of the rater*. For instance, items # 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* reflected the participants' attitudes towards holistic and

analytic assessment. As different aspects of speaking performance were focused simultaneously during analytic assessment and the final score was 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 had thought. For example, one of the instructors wrote that since he was supposed to give only one score during holistic assessment, the score was a real reflection of his impression. However, when he was asked to use the analytic assessment tool, he had to give separate scores and then add them. The final score, he said, was generally higher or lower than he had anticipated.

Another theme observed was *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* were highly related to this theme. The most common reason given for *practicality* was the ease holistic assessment provided for the assessment of the whole. When the work load of instructors is considered, practicality is a big issue. Therefore, it is quite understandable that instructors' in the present study showed a positive attitude towards this aspect of holistic assessment in terms of *practicality*. *The time needed for assessment* was another theme that was seen in the attitudes of the participants in respect with holistic and analytic assessment. This theme matched with the items # 3, *Holistic assessment is time consuming* and # 12, *Analytic assessment takes a long time*. In fact, a reverse coded item, # 3 had the highest mean score, which meant that there was a strong agreement within the participants with the statement and # 12 received the least positive score. One of the instructors wrote that trying to score different aspects at the same time was a hard thing to do. Another one wrote, "instead

of one score, you try to give five or six scores, which makes it harder and harder”. As a result, the participants seemed to pay attention to the time assessment takes and that was another reason for them to have a positive attitude towards holistic assessment.

The freedom allowed by the assessment tool was another theme that emerged from the responses of the participants, which turned out to be in favor of the holistic assessment. This theme matched with the items # 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 wrote that he did not like being controlled and trying to make his scores fit in to the criteria and that was why he preferred holistic assessment. When the mean scores for each item for this theme is looked at, the difference in the mean score can lead people to consider that the participants thought holistic assessment allowed them to score with more freedom. *Assessing speaking holistically* was another theme supported by items # 9, *Speaking performance should be assessed as a holistically* and # 18, *Speaking performance should be assessed analytically*. Nine out of ten instructors who gave this as one of the reasons for their positive attitudes towards holistic assessment explained that the whole was 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 towards holistic assessment.

Yet, the most common theme in favor of analytic assessment seen was the *feedback opportunities it created*. Not surprisingly, item # 16, *Better feedback for students can be provided with holistic assessment* received the least positive attitude score, indicating that on average participants *disagreed* with the statement while the instructors seemed to *strongly agree* on item # 7, *Analytic assessment can provide better feedback for students*. The instructors considered analytic assessment as a more useful type of assessment to provide feedback for students to improve their

skills. For instance, one of them wrote “at the end of analytic assessment, I know what my students need to do to develop themselves”. Finally, *the comfort analytic assessment created* was another theme emerged in favor of the analytic assessment. Item # 15, *I feel more confident and comfortable when assessing performance with a holistic assessment tool* had 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 made her job easier since following guidelines is easy. Yet, this element was only seen in the responses of the participants to the open-ended items in favor of analytic assessment and was not seen for holistic assessment.

Third, there were also some themes that highlighted different issues than those mentioned in the items of the attitude questionnaire. For example, the most common criticism towards holistic assessment was the *subjectivity of holistic assessment*. The six instructors who criticized holistic assessment for being subjective suggested that stricter guidelines were required for more reliable scores. For example, one of them wrote “when I use analytic rubric, it guides me and the scores I assign become more reliable”. *The reliability of the scores* obtained by analytic assessment was another point made by instructors. The instructors seemed to believe that analytic assessment provided more reliable scores than those provided by holistic assessment. For example, one instructor wrote that “in order to reach more reliable scores, guides are needed and analytic scoring provides this”. Parallel with that, *explaining the rating process* was another common criticism for holistic assessment. The four instructors wrote about their concern for the *vagueness of the rating process* when holistic assessment was used. In fact, one of them criticized it as a tool lacking concrete guidelines, which, inevitably, led to unreliable scores. Another criticism for holistic assessment was related to rater training. The instructors pointed 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 the attention*, that came up only in the responses of the participants was about the analytic assessment and the cognitive load it brought. In fact, one instructor wrote that, he could not follow the student when he had to focus on different aspects of his/her performance and this could lead to a failure in the assessment because he had 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 which instructors pay attention to. First of all, it seems like ease of use and time were among the major issues the instructors consider when it comes to assessment. They seemed to consider these two as critical and had a tendency to favor the type of assessment which enables them to assess easily and quickly. This might be one of the reasons for the positive attitude towards holistic assessment. Secondly, the instructors in the study seemed to believe that speaking should be assessed as a whole rather than from different aspects or components. As mentioned above, they had a tendency to consider speaking skill as the final performance which included 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 were 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 a performance separately, the final score obtained might end up different from what was 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 own impressions. Finally, *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 seemed to be an issue with the holistic assessment as one final score to cover all components was thought to lead to unexplainable scores.

When these findings were considered together, it was seen that the attitudes of the instructors seemed 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 holistic assessment. In fact, as Luoma (2004) highlighted, this might alone be one the important reasons for instructors to choose holistic assessment over analytic assessment. Similarly, Huot (1990) highlighted *flexibility* or *more freedom* as advantages of holistic assessment and this was seen as another advantage the instructors in this study mentioned. On the other hand, as Brown (2004) and Weigle (2002) pointed, holistic assessment offered no feedback or had no washback effect for students but analytic assessment did. This was the strongest criticism for holistic assessment and was put forward for the analytic assessment by the instructors. The instructors' other criticism for the holistic assessment was related to the impossibility to monitor the assessment process. According Bachman and Palmer (1996), during holistic assessment, monitoring the assessment procedure was very difficult and the instructors in the study seemed to have noticed it as well. Also, as Brown (2004) and Luoma (2004) emphasized, the subjectivity of the scores was one of the most common criticism 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) were among the reasons for the instructors to support analytic assessment of speaking. Also, feeling more comfortable because of having a more detailed scoring tool (Knoch, 2009) was another reason for the positive attitude towards 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) were among the criticism put forward by the participants in this study for analytic assessment.

Research Question 2: Is There a Difference in the Speaking Exam Scores Assigned by Instructors Using Holistic and Analytic Assessment Tools?

The findings suggested that regardless of the tool used to assess speaking performance in this study, the scores assigned by the instructors did not show any statistically significant difference. This finding contradicted with Carr (2000) and Barkaoui (2010), who reported that the type of assessment tool used had an effect on the scores. However, the findings of this study were parallel with those of Chuang (2009), who found 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 homogeneity of the speaking samples might have had an effect on the results. Another reason might be related to the instructors, which might be called as decision-making behaviors of the instructors (Barkaoui, 2007). Although the instructors were asked to assess the speaking exams analytically, they might have ignored the components of the tools, assigned an overall score and then divided them to components of the assessment tool.

Research Question 3: Do the Speaking Exam Scores Obtained via Holistic and Analytic Assessment Differ According to the Instructors' Background?

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

Age and Scores

The assigned speaking scores were compared by age groups of the participants by employing One Way ANOVA and it was found the scores did not differ in terms of age groups. The findings were different from the findings of Chuang (2009), which stated that younger teachers were more severe in terms of scoring while the older teachers were more lenient. The reason for the non significant difference in this study might have been caused by the low number of the participants. For example, if the sample size was bigger, with more participants with different ages, this might have resulted 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.

Gender and Scores

As for the comparison of the speaking exam scores by the gender of the participants, an independent samples *t*-test was used and it was seen that the scores did 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) reported that male teachers scored higher than female teachers, Gholami et al. (2011) and O'Sullivan and Porter (1996) reported 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 found no statistically significant difference in terms of the scores of male and female teachers, which clearly showed that gender of the instructors did not lead to a difference in the scores.

Academic Major and Scores

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

Years of Experience and Scores

The comparison of the speaking exam scores obtained from holistic and analytic assessment tools with years of experiences of the instructors by employing One Way ANOVA showed interesting results. For instance, although a statistically significant difference was not seen in terms of analytic assessment, there was a statistically significant difference among participants' years of experience and the

scores they assigned using holistic assessment tool. Specifically, the less experienced instructors were 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 instructors' were more lenient seemed to be different from the studies reported in the literature. For instance, although Chuang (2009) indicated no statistically significant difference between years of experience, Song and Caruso (1996) found that raters with more years of experience had a tendency to be more lenient when using holistic assessment tools. Similarly, Huang and Jun (2015) reported that inexperienced raters were 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 instructors' may be less familiar with the assessment tool and/or assessing speaking. If a longitudinal research were conducted, this finding might be confirmed or rejected.

The fact that such a difference in scores was 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 seemed to assign similar scores but when they were more free and without stricter guidelines, the less experienced instructors seemed to score more freely. As a result, such a difference in the scores were seen.

Pedagogical Implications

This study had 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 suggested that instructors in this study had different attitudes towards holistic and analytic assessment, which were based on their ideas about

assessment of speaking. For example, the instructors who had positive attitude for the holistic assessment seemed to reject piece by piece assessment of speaking and vice versa. Yet, the findings also showed that instructors had a more positive attitude towards holistic assessment. Secondly, the findings revealed that there was not a statistically significant difference between the scores using the holistic and analytic assessment instruments.

Even though more data is required, yet, within the 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 were close to each other in this study and instructors displayed varying attitudes towards holistic and analytic assessment considering advantages and disadvantages of both types of assessment, allowing them to choose the tool to use may help them to perform better during assessment. As instructors have a preference for the type of assessment tool they use, using the assessment they prefer this might lead to more satisfaction for the instructors and the institution.

Another pedagogical implication of the study is related to using different types of assessment tools simultaneously. Though this was not focused in 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.

Limitations of the Study

This study was conducted at a state university in Turkey. The number of the instructors who took part in this study was twenty four and the number of the students, whose speaking exams were assessed, was ten. Because of administrative and feasibility issues, the number of the instructors and the students had be kept

relatively small. Also, the instructors who participated in this study had at least six years of experience in ELT. This was because of the fact that EU SFL had previously used both types of assessment for speaking. Since these instructors were the ones who had experienced using both of them, it was assumed that they would have been more aware of the differences between both types of assessment.

Before the beginning of the study, the original plan was to include eight more students in the study as decoys to minimize the recall effect for the students' performances. In other words, originally, eighteen students were planned to take part in the study, fourteen for the holistic assessment and fourteen for the analytic assessment. Of the eighteen, only ten students were to have been assessed using both the holistic and analytic assessment tools. However, because of administrative issues, only ten students participated in the study. As the interval between the holistic and analytic assessment was a month, this was thought to be sufficient to prevent the recall effect. Nevertheless, the results of the study should be interpreted with this in mind.

Suggestions for Further Research

Firstly, related to the pedagogical implications presented in the previous section, another study may focus on whether instructors' preferences regarding the assessment tools affect the scores they assign. In other words, allowing instructors to make a choice between using 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 were included in the study. In other words, novice instructors did 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 were only asked to assess the sample speaking exams silently. However, in a future study, 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 was pre-intermediate since there were only pre-intermediate learners at EU SFL by the time this study was conducted. 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.

Conclusion

This study aimed to find out language instructors' attitudes towards holistic and analytic assessment and investigating whether the scores assigned by using each type of assessment differed or not. The study also examined whether, and if any of, these differences in scores were related to instructors' background. The study was conducted at the School of Foreign Languages, Erciyes University with twenty four language instructors and the data were gathered through an attitude questionnaire and holistic and analytic assessment tools. The findings showed that the instructors in this study had varying attitudes towards holistic and analytic assessment, more in favor of the holistic assessment. These attitudes revolved around advantages and disadvantages of both types of assessment. For instance, practicality of the tool, reliability of the scores, ease of use and time were among the common themes emerged. It was also found that the scores assigned by holistic and analytic assessment tools did not differ and except for years of experience and speaking exam

scores obtained by holistic assessment, there were no difference in the scores assigned by holistic and analytic assessment tools and the background of the instructors. In the light of this study's findings, it can be suggested that since the scores of speaking assessment did not differ regardless of the assessment tool being used, allowing instructors to choose between holistic and analytic assessment tools may be considered for further educational purposes.

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 Press.
- 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.
- 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 (Eds.), *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 influence of analytic versus holistic thinking. *Journal of Consumer Research*, 33, 529-536.
- 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.
- O'Sullivan, B., & Porter, D. (1996). Speech style, gender, and oral proficiency interview performance. *Paper presented at the Annual Meeting 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>

- Plass, J. A., & Hill, K. T. (1986). Children's achievement strategies 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.
- 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 - 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 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 - Sample Transcription of a 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

...