THE EFFECT OF RATERS' PRIOR KNOWLEDGE OF STUDENTS' PROFICIENCY LEVELS ON THEIR ASSESSMENT DURING ORAL INTERVIEWS

A MASTER'S THESIS

BY

FATMA TANRIVERDİ-KÖKSAL

THE PROGRAM OF
TEACHING ENGLISH AS A FOREIGN LANGUAGE
BİLKENT UNIVERSITY

ANKARA

SEPTEMBER 2013

The Effect of Raters' Prior Knowledge of Students' Proficiency Levels on Their Assessment During Oral Interviews

The Graduate School of Education
of
Bilkent University

by

Fatma Tanrıverdi-Köksal

In Partial Fulfillment of the Requirements for the Degree of Master of Arts

in

The Program of
Teaching English as a Foreign Language
Bilkent University
Ankara

September 2013

To my beloved husband, Kerem Köksal,

&

To my family.

BİLKENT UNIVERSITY
THE GRADUATE SCHOOL OF EDUCATION
MA THESIS EXAMINATION RESULT FORM

23 September, 2013

The examining committee appointed by The Graduate School of Education for the
Thesis examination of the MA TEFL student
Fatma Tanrıverdi-Köksal
has read the thesis of the student.
The committee has decided that the thesis of the student is satisfactory.

Thesis Title:	The Effect of Raters' Prior Knowledge of Students' Proficiency Levels on Their Assessment During Oral Interviews
Thesis Advisor:	Dr. Deniz Ortaçtepe Bilkent University, MA TEFL Program
Committee Members:	Asst. Prof. Dr. Louisa Buckingham Bilkent University, MA TEFL Program
	Asst. Prof. Dr. Zeynep Koçoğlu Yeditepe University, Faculty of Education

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 Teaching English as a Foreign Language.

____
(Dr. Deniz Ortaçtepe)
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 Teaching English as a Foreign Language.

____
(Asst. Prof. Dr. Louisa Buckingham)
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 Teaching English as a Foreign Language.

____
(Asst. Prof. Dr. Zeynep Koçoğlu)
Examining Committee Member

Approval of the Graduate School of Education

____
(Prof. Dr. Margaret Sands)
Director

ABSTRACT

THE EFFECT OF RATERS’ PRIOR KNOWLEDGE OF STUDENTS’ PROFICIENCY LEVELS ON THEIR ASSESSMENT DURING ORAL INTERVIEWS

Fatma Tanrıverdi-Köksal

M.A. Department of Teaching English as a Foreign Language
Supervisor: Dr. Deniz Ortaçtepe

September, 2013

This quasi-experimental study, focusing on scorer reliability in oral interview assessments, aims to investigate the possible existence of rater bias and the effect(s), if any, of the raters’ prior knowledge of students’ proficiency levels on rater scorings. With this aim, the study was carried out in two sessions as pre and post-test with 15 English as a foreign language (EFL) instructors who also perform as raters in the oral assessments at a Turkish state university where the study was conducted.

The researcher selected six videos as rating materials recorded during 2011-2012 academic year proficiency exam at the same university. Each of these videos included the oral interview performances of two students. The data collection started with a norming session in which the scores the raters assigned for the performances of four students recorded in two extra videos were discussed for standardization. After the norming session, using an analytic rubric, the participants performed individually as raters in the pre and post-test between which there was at least five week interval. In both the pre and post-test, the raters were asked to provide verbal reports about what they thought while assigning scores to these 12 students from three different proficiency levels. While no information about students’ proficiency levels were provided to the raters in the pre-test, the raters were informed about students' levels both in oral and written format in the post-test. The scores the raters assigned were filed, and the think-alouds were video-recorded for data analysis.

As a result, quantitative data analysis from the pre and post-test scores indicated that there was a statistically significant difference between the pre and post-test scorings of eight raters assigned to different components of the rubric such as *Vocabulary, Comprehension,* or *Total Scores* which represented the final score each student received. Further analysis on all the *Total Scores* assigned for individual students by these 15 raters revealed that compared to pre-test scores, ranging from one point difference to more than 10 points, 75 % of the *Total Scores* assigned by these raters ranked lower or higher in the post-test while 25 % did not change. When all the raters' verbal reports were thematically analyzed in relation to the scores they assigned and the references they made to the students' proficiency levels, it was observed that 11 raters referred to the proficiency levels of the students while assigning scores in the post-test. Furthermore, the *Total Scores* assigned for each group of students each of which consisted from a different proficiency level were analyzed, and the results indicated that the raters differed in their degree of severity/leniency while assigning scores for lower and higher level students.

Key words: rater effects, rater bias, rater/scorer reliability, intra-rater reliability, oral interviews, oral assessment, think-aloud protocols.

ÖZET

NOTLANDIRANLARIN ÖĞRENCİLERİN DİL YETERLİLİK SEVİYESİNİ ÖNCEDEN BİLİYOR OLMASININ ONLARIN SÖZLÜ MÜLAKAT ESNASINDAKİ NOTLARDIRMALARINA ETKİSİ

Fatma Tanrıverdi-Köksal

Yüksek Lisans, Yabancı Dil Olarak İngilizce Öğretimi Bölümü
Tez Yöneticisi: Dr. Deniz Ortaçtepe

Eylül 2013

Bu yarı deneysel çalışma, sözlü mülakatların değerlendirilmesinde notlandırıcı güvenirliğine odaklanarak, olası notlandırıcı önyargısını ve notlandıranların öğrencilerin dil yeterlilik seviyelerini önceden biliyor olmasının verdikleri notlar üzerinde var ise etkilerini araştırmayı amaçlamaktadır. Bu amaç doğrultusunda bu çalışma, çalışmanın uygulandığı Türkiye'deki bir devlet üniversitesinde yabancı dil olarak İngilizce öğreten ve aynı üniversitede sözlü sınavlarda notlandırıcı olarak görev alan 15 okutman ile ön ve son test olarak iki oturumda yürütülmüştür.

Araştırmacı, aynı üniversitede 2011-2012 akademik yılı muafiyet sınavı esnasında kaydedilmiş altı videoyu notlandırma materyeli olarak seçmiştir. Bu videoların her biri iki öğrencinin sözlü performansını içermektedir. Veri toplama, notlandıranların iki ekstra videoda kayıtlı dört öğrencinin performansına verdikleri notların standardizasyon için tartışıldığı norm belirleme oturumu ile başlamıştır. Norm belirleme oturumundan sonra, katılımcılar analitik bir kriter kullanarak arasında en az beş hafta olan ön test ve son testte bireysel olarak notlandırıcı görevini üstlenmişlerdir. Hem ön hem de son testte, notlandırıcılardan üç farklı seviyeden bu 12 öğrenci için not verirken ne düşündükleri ile ilgili sözlü bildirimde bulunmaları istenmiştir. Öğrencilerin dil yeterlilik seviyeleri ile ilgili ön testte herhangi bir bilgi verilmezken, notlandıranlar öğrencilerin seviyeleri konusunda son testte sözlü ve yazılı olarak bilgilendirilmiştir. Veri analizi için notlandıranların verdikleri notlar dosyalanmış, sesli-düşünme protokolleri video kaydına alınmıştır.

Sonuç olarak, ön ve son test notlarının nicel veri analizi, sekiz notlandırıcının kriterin *Kelime, Anlama,* ya da her öğrencinin aldığı son notu temsil eden *Toplam Not* gibi farklı bölümlerinde verdikleri ön ve son test notları arasında istatistiksel olarak anlamlı bir fark olduğunu göstermiştir. 15 notlandırıcı tarafından her bir öğrenci için verilen *Toplam Notların* daha detaylı incelenmesi, ön test notlarına kıyasla, notlandırıcılar tarafından verilen *Toplam Notların* % 75'inin, bir puandan 10 puandan fazlaya kadar çeşitlilik göstererek, son testte düştüğü veya yükseldiği, fakat % 25'inin değişmediği saptanmıştır. Tüm notlandıranların sözlü bildirimleri, verdikleri notlar ve öğrencilerin dil yeterlilik seviyelerine değinmeleri ile bağlantılı tematik olarak incelendiğinde, 11 notlandıranın son testte not verirken öğrencilerin dil yeterlilik seviyelerine değindikleri gözlemlenmiştir. Ayrıca, her biri farklı bir dil yeterlilik seviyesinden oluşan her bir öğrenci grubu için verilmiş *Toplam Notlar* incelenmiş ve sonuçlar notlandıranların düşük veya yüksek dil yeterlilik seviyesi öğrencileri için not verirken, hoşgörü ve katılık derecesi açısından farklılık gösterdiğini ortaya çıkarmıştır.

Anahtar kelimeler: notlandıran etkisi, notlandıran güvenirliği, tek notlandıran güvenirliği, sözlü mülakatlar, sözlü notlandırma, sesli düşünme protokolleri.

ACKNOWLEDGEMENTS

My deepest gratitude and heartfelt appreciation should go to several important figures in my life without whom I would not be able to finish this final “product of achievement.”

First and foremost, I would like to express my deepest gratitude to my supervisor Dr. Deniz Ortaçtepe for her continuous and invaluable support, patience and encouragement. Whenever I needed her wisdom, she was there with her diligence, constructive feedback, support, and encouragement. I would not have been able to complete my study without her. I am also deeply indebted to her for having faith in my potential to be able to conduct such an intensive, challenging study. Once again, I want to thank her for becoming my mentor and enlightening me with her guidance in this challenging journey.

Second, I am grateful to Asst. Prof. Dr. Julie Mathews-Aydınlı for her valuable suggestions and constructive feedback. Thanks to her insightful and thought-provoking ideas at the very early stages of my study, I have created such a piece of work. I further would like to express my gratitude to my committee members, Asst. Prof. Dr. Louisa Buckingham and Asst. Prof. Dr. Zeynep Koçoğlu for their insightful comments, precious feedback, and support.

I also wish to extend my gratitude to my institution, Bülent Ecevit University, Prof. Dr. Mahmut Özer, the President, and Prof. Dr. Muhlis Bağdigen, the Vice President, for giving me the permission to attend this well-respected and highly acknowledged MA TEFL program and providing me the opportunity to take part in such a privileged program. I am also very grateful to my Director, Inst. Okşan Dağlı, for giving me the permission to attend Bilkent MA TEFL and being understanding and supportive during this demanding journey.

I owe my gratitude to all the teachers I have had so far, especially Mustafa Özdemir, my primary school teacher and Ayşe Pınar İnsel, my high school English teacher for changing my destiny by being there.

In addition, I owe many thanks to every member of the MA TEFL Class of 2013, best class ever. Thanks to this unique group, this year has become an unforgettable experience for me. Many thanks go to especially Selin Yıldırım and Işıl Ergin for their never ending wake-up calls. Also, Dilara Yetkin and Selin Yıldırım, you are my "*better late than never*."

Special thanks to all my participants for the time and energy they put in this study. Without them, it would not be possible for me to conduct this research. I will always be indebted to each of them for their support.

Thank God I have several friends with whom I am stronger. My lifelong friend, "*Kankam*", Ayça, with whom I have lived half of my life, who was, has always been, and will be with me, and her dear husband Alper deserve special thanks for the good weekends in Ankara. I would also like to thank to my dear colleagues at Anadolu, especially dear friends Tuğba, Ayça, Hülya Hocam, and Müge Hocam, for their constant support, messages, and phone calls, and Özlem, Aslı, and Meral, for their friendship and especially for their support when I decided to apply for this program. I should also thank to my dear friends/sisters Füsun, Funda, Sibel, and Göze for their precious friendship. Last but not least, my dear Eda should receive my gratitude for always being with me and understanding me before I say anything.

I am particularly grateful to my family members who have always showed their faith in me. My special and sincere thanks go to my parents Hasan and Halime Tanrıverdi for always being supportive whenever I need them.

Finally, my deepest and greatest gratitudes go to my beloved husband, Kerem

Köksal, “*Rosicky’m*,” for his constant support, encouragement and unconditional love. We have grown up together, faced and overcome all the challenges hand in hand. Without him, I would not have become the person I am now.

Thank you all for somehow being in my life.

TABLE OF CONTENTS

LIST OF TABLES

Table

[1] The raters who had significant differences between their pre and post-test scorings.

[2] The raters, Rater # 2, Rater # 7, Rater # 9, Rater # 10, and Rater # 12, without a significant difference but with reference to the levels

[3] The raters, Rater # 5 and Rater # 13, without a significant difference and no reference to the levels

LIST OF FIGURES

Figure

[4] The raters, Rater # 2, Rater # 7, Rater # 9, Rater # 10, and Rater # 12, without a significant difference but with reference to the levels

[5] The raters, Rater # 5 and Rater # 13, without a significant difference and no reference to the levels

CHAPTER I: INTRODUCTION

Introduction

Teaching and testing, which are two key entities of education, cannot be considered as distinct and independent from each other (Rudman, 1989) because when there is teaching, it is usually accompanied by testing to examine to what extent the learners have acquired the desired learning outcomes. With the growing popularity of the communicative theories of language teaching in the 1970s and 1980s (Brown, 2004; McNamara, 1996), more traditional test formats such as pencil-and-paper tests have been replaced by communicative approaches to language learning, teaching, and testing which introduced performance assessment as an alternative assessment instrument that focuses on what learners can do with the language (McNamara, 1996). In other words, rather than answering questions that require limited response and focus mostly on receptive skills, the learners are expected to demonstrate command of productive skills by performing the given tasks effectively. Once the importance of assessing communicative competency has been acknowledged, oral interviews have taken its place in academic contexts as one of the alternative assessment instruments to evaluate students' spoken proficiency. However, although widely conducted, there has been an ongoing debate on the reliability of oral interview resulting scores due to the existence of human raters and the differences in their scorings.

Several studies conducted on rater effects have revealed that human raters vary in their scoring behaviors because of several factors such as their educational and professional experience, nationality and native language, rater training, and candidates' and/or inteviewers' gender (e.g., Chalhoub-Deville, 1995; Chalhoub-Deville & Wigglesworth, 2005; Galloway, 1980; Lumley & McNamara, 1995;

O'Loughlin, 2002; O'Sullivan, 2000, 2002; Winke & Gass, 2012; Winke, Gass, & Myford, 2011), but the factors that affect raters' behaviors, scoring process, and final scorings in oral interviews have not been completely explored (Stoynoff, 2012). The fact that such factors can lead to misinterpretations and misjudgments of test-takers' actual performances, and thus, affect their academic success and future has generated the need to further explore these construct-irrelevant factors. However, there is a limited body of research focusing on cognitive processing models, especially verbal reports of raters, to investigate how raters assign scores in oral interviews and provide better insights into the raters' decision making process. For this reason, with the help of this study, it is hoped to contribute to the existing literature by revealing another source of rater effects, and thus be of benefit to the test-takers, raters, and institutions.

Background of the Study

Current language teaching approaches, including Communicative Language Teaching (CLT), brought about new alternative assessment instruments to language testing, one of which is oral interviews (Caban, 2003; Jacobs & Farrell, 2003). Oral interviews are widely used in proficiency tests which are conducted for different purposes such as to determine whether learners can be considered proficient in the language or whether they are proficient enough to follow a course at a university (Hughes, 2003). Oral interviews are usually conducted in three formats; individually, in pairs, and in groups; and single or two interlocutors and/or raters usually evaluate the performance of learners during interviews.

Although oral interviews are widely used in academic contexts, they are still considered as a controversial type of assessment. One of the main concerns related to oral interviews is that some degree of subjectivity is likely to affect the ratings because human raters are the ones determining the scores during oral interviews (Caban, 2003). Because testing of spoken language to assess communicative competence is open to raters’ interpretations (e.g., interpretation and/or application of the scoring criteria, Bachman, 1990) and rating differences (Ellis, Johnson, & Papajohn, 2002), concerns about validity and reliability, which are two important qualities of a test (Bachman & Palmer, 1996), have been the center of the discussion of oral interviews for a long time (Joughin, 1998). While validity refers to whether a test is measuring what it is supposed to measure (Hughes, 2003), reliability refers to “the consistency of measurement” (Bachman & Palmer, 1996, p. 19), that is, no matter when or where they take it, the test-takers will receive similar scores (Brown, 2004). However, Bachman (1990) suggests that instead of taking them as two different aspects of measurement, they should be considered together in order to understand and control the factors that may affect test scores.

Reliability, for which Weir (2005) uses the term “scoring validity” (p. 22), is the focus of this study. While there are different types of reliability, rater reliability, which is the focus of this study, is a term used to refer to the consistency of the raters in their scorings (Weir, 2005). Since the existence of human rater has been acknowledged as one of the many challenging factors that can change a score assigned to a test performance (Hardacre & Carris, 2010), Hughes (2003) points out that when the decision or the result is very important for the test takers as it is in high stakes exams, achieving high reliability also becomes very important.

Hence lower rater reliability affects the raters’ scorings negatively and causes detrimental effects for the test-takers such as failure in the exam and lower academic success, it contradicts with another important quality of a test: fairness in testing, which can only be assured by providing equal opportunities to candidates considering test design, test conduct, and scoring (Willingham & Cole, 1997). The existence of differences in rater behaviors in terms of more lenient or severe rating than what learners' actual performance should receive has led researchers to look at another aspect of fair scoring: bias which is an important concept in language testing since test results should be "free from bias" (Weir, 2005, p. 23). McNamara and Roever (2006) define bias as "a general description of a situation in which construct-irrelevant group characteristics influence scores" (p. 83). In other words, bias in assessment refers to an unfair attitude toward one side by favoring or disadvantaging one or some test takers. As a result, lower reliability and the existence of rater bias in oral interviews, as well as in other forms of assessment, can highly affect the decisions about the test-takers' performances and lead to raters' misjudgments about the test-takers' performances, and thus, prevent the raters from assigning fair and objective test results.

In the literature, rater effect, rater error, rater variation, and rater bias usually refer to the same issue: the change in rater behaviors affected by factors other than the actual performance of test-takers. As Fulcher and Davidson (2007) state, several studies have been conducted to find out how personal and contextual factors affect interlocutors' and raters' behaviors and decisions, and how these factors can be controlled to eliminate or limit the human rater factor in scores. Previous studies have investigated rater effect on oral test scores from different perspectives such as the effects of raters' educational and professional experience (e.g., Chalhoub-Deville, 1995; Galloway, 1980), the effects of raters' nationality and native language (e.g., Chalhoub-Deville & Wigglesworth, 2005; Winke & Gass, 2012; Winke et al., 2011), the effects of rater training (e.g., Lumley & McNamara, 1995; Myford & Wolfe, 2000), and the effects of candidates' and/or inteviewers' gender (e.g., O'Loughlin, 2002; O’Sullivan, 2000, 2002). A great deal of the studies which investigated the rater effect on oral test scores have revealed that beliefs, perceptions and bias of raters are important factors that can affect the test results.

McNamara (1996) points out that “judgments that are worthwhile will inevitably be complex and involve acts of interpretation on the part of the rater, and thus be subject to disagreement.” (p. 117). Joe (2008) also emphasizes the complex procedural and cognitive process the raters go through while assigning scores in performance assessments. He suggests that human scoring involves two important principles “what raters perceive and how raters think” (Joe, 2008, p. 4). For this reason, due to the fact that statistical approaches fail in providing a full understanding of the decision making process, recent studies have started to show interest in applying cognitive processing models in order to gain better insights into how raters assign scores, and why there are differences among raters’ scorings. However, since they have been used only in recent studies conducted on rater effects in oral interviews, there is a limited body of research focusing on these models in oral interview assessment. A frequently used qualitative data collection method for exploring cognitive processes of raters, verbal report analysis has two types: (a) concurrent verbal reports, also referred to as think-alouds, are conducted simultaneously with the task to be performed, and (b) retrospective verbal reports are gathered right after the performance task (Ericson & Simon, 1980). Think-aloud protocols are considered as more effective in understanding raters’ cognitive processing during oral assessment scoring because it is sometimes difficult to remember what someone did and why he/she did it (Van Someren, Barnard, & Sandberg, 1994). For this reason, while investigating the rater effects in oral assessment, employing think-aloud protocols for understanding what raters think and how they assign a score can be very effective. As Fulcher and Davidson (2007) suggest, in oral assessments, for which subjective scoring of human raters is at the center of debate, the attempts to control the construct-irrelevant factors, the factors other than the actual performances of test-takers, are crucial in order to provide and guarantee fairness in large-scale testing.

Statement of the Problem

In many countries, oral interviews are widely used in academic contexts for the purpose of measuring oral language proficiency although it has been acknowledged that rater factors have a considerable effect on the differences in resulting test scores (Lumley & McNamara, 1995). Due to the ongoing debate on the reliability of oral assessment scorings, several researchers have investigated whether some external factors have an effect on raters' scoring process and final test results. For instance, Lumley and McNamara (1995) examined the effect of rater training on the stability of rater characteristics and rater bias whereas MacIntyre, Noels, and Clément (1997) investigated bias in self-ratings in terms of participants' perceived competence in an L2 in relation with their actual competence and language anxiety. O'Loughlin (2002) and O' Sullivan (2000) looked into the impact of gender in oral proficiency testing while Caban (2003) examined whether raters' language background and educational training affect their assessments. Chalhoub-Deville and Wigglesworth (2005) investigated if raters from different English speaking countries had a shared perception of speaking proficiency while Carey, Mannell, and Dunn (2011) studied the effect of rater's familiarity with a candidate's pronunciation. Although there are several studies conducted on various rater effects on oral performance assessment, defining the factors that affect rater judgment is still in the exploratory stage; and to the knowledge of the researcher, no study has been conducted to investigate rater effects in oral interviews in terms of the effect that raters' prior knowledge of students' proficiency levels may have on their assessment behaviors.

In Turkey, oral interviews are widely used in university preparatory schools-intensive English programs as an alternative assessment in midterms and final exams. Although rubrics are always used, raters may behave differently both in their own scoring processes and from each other while conducting the interviews, interacting with the test-takers and assessing the test-takers' performances. As a result, in many cases, neither the test-takers nor the classroom teachers are content with the results because if raters are affected by some factors other than the actual performances of test-takers during the rating process, it is highly possible that they can misjudge the performance of test-takers which can lead to the misinterpretation of scores (Winke et al., 2011). In other words, due to the rater measurement error which results from the effects of some performance-irrelevant factors, a student can get a lower score than he/she deserves, or even worse, fail in the test. For this reason, the institutions and/or the raters are sometimes sued by the test-takers due to the fact that oral interviews are high-stakes exams in terms of their critical effects on the decisions for students' pass or fail scores. Considering the fact that human raters may sometimes yield to subjectivity in their ratings (Caban, 2003), investigating rater effects in oral interview scores is of great importance for accurate assessments because the results of inaccurate judgments may have harmful effects for test-takers, raters, and the institutions. Therefore the present study will investigate the following research question:

- To what extent does raters' prior knowledge of students' proficiency levels influence their assessment behaviors during oral interviews?

Significance of the Study

Since oral interviews are assessed by human raters, it is almost inevitable that some raters will behave differently in their scorings, especially, if they are affected by some construct-irrelevant factors. If rater effects exist in the scorings, it jeopardizes the reliability and fairness of a test. Considering rater measurement error as a very influential negative impact on test-takers' academic achievement, any attempt to diminish the effects of external factors such as rater effect is noteworthy. However, using merely statistical approaches to explore rater effects in oral assessment cannot provide significant information about what and how raters think while scoring in oral assessment procedures. This mixed-method study, using both statistical approach and verbal reports of raters, may augment the existing literature on rater bias in oral assessment by showing any possible effects of the raters' prior knowledge of students' proficiency levels on their scorings, and thus, by revealing another form of rater bias.

At the local level, during oral interviews, it is sometimes observed that the comments of raters on the test-takers' performance sometimes provide evidence of different types of bias such as the effects of the accent, the anxiety level, and the physical appearance of the test-takers on raters' scorings. Moreover, due to the subjective scorings of human raters, the relatively high differences in scores may sometimes cause the institutions, teachers, and students to question the reliability of oral interviews as a type of assessment; moreover, some may also argue for abandoning oral assessment at all although in current approaches to teaching, it is crucial to teach and assess speaking skill. Thus, the results of this study may be of benefit to test-takers, raters and administrators by providing better insights into how raters assign their scores. Moreover, raising awareness about the possible existence of a different type of rater effects may prevent rater judgmental errors and further arguments about the reliability of oral interviews; and by doing so, the goal of ensuring fair tests can also be achieved.

Conclusion

In this chapter, a synopsis of the literature on performance assessment in oral interviews and concerns about subjective scoring has been provided through a brief introduction of key terms, the statement of the problem, research question, and the significance of the study. The next chapter will review the relevant literature on language testing, assessment of speaking ability, factors that affect speaking assessment, and existing measurement approaches to test rater reliability.

CHAPTER II: LITERATURE REVIEW

Introduction

The aim of this chapter is to introduce and review the literature related to this study investigating the possible effects of raters' prior knowledge of students' proficiency levels on their assessment behaviors during oral interviews in proficiency exams. In the first section, language testing in relation to types of tests will be covered with a particular focus on proficiency tests. A brief introduction of qualities of a test will also be provided in this section, especially focusing on the issues of reliability and fairness in testing. In the second section, literature on the assessment of speaking ability will be reviewed in relation to formats of speaking tests, especially oral interviews. In the next section, factors that affect speaking assessment will be elaborated with an extensive focus on rater related factors and rater effects on test scores. In the last section, current research about the existing measurement approaches on rater effects will be covered. This part will continue with a detailed discussion of verbal report protocols, especially Think-Alouds.

Language Testing

According to Brown (2004), a test is "a method of measuring a person's ability, knowledge, or performance in a given domain" (p. 3). In other words, tests are used to measure what a person knows about a specific topic, and what he/she can do with that knowledge. Similarly, language tests are used to assess people's knowledge and performance in that language, and they are used for several purposes such as to determine whether learners can be considered proficient in the language or whether they are proficient enough to follow a course at a university (Hughes, 2003). There exist several types of tests depending on their purpose.

Types of Tests

The following section focuses on the most common used types of tests in educational settings, which are classified into four categories according to the purpose of their use and types of information they provide (Hughes, 2003). The four types of tests which will be discussed in this section are achievement tests, diagnostic tests, placement tests, and proficiency tests.

Achievement tests.

Achievement tests are used to make decisions about how much the learners have learned within the program (Brown, 1996). They are used to find out how much the students have achieved the desired learning outcomes of the course and the program (Hughes, 2003). They are also used to evaluate the effectiveness of the teaching and the language programs (Bachman & Palmer, 1996). Thus, they are "associated with the process of instruction" (McNamara, 2000, p. 6), and are administered during or at the end of a course.

Diagnostic tests.

Diagnostic tests are used to assess the strengths and weaknesses of learners (Brown, 1996; Hughes, 2003) "for the purpose of correcting an individual's deficiencies "before it is too late"" (Brown, 1996, p. 14, emphasis in original). These tests are used to make decisions about the problems a learner may have in his/her learning process. In other words, diagnostic tests are designed to determine the specific problematic areas at which learners have difficulty in achieving the learning outcomes of the course.

Placement tests.

Placement tests are used to place the students at the classes that are appropriate to their language proficiency (Hughes, 2003). They are conducted at the beginning of a course to group students with similar language ability and organize homogenous classes so that lessons and curriculum can be planned according to the learning points appropriate for that level of students (Brown, 1996).

Proficiency tests.

The last type of test to be discussed is proficiency tests. According to Longman Dictionary of Language Teaching and Applied Linguistics (LTAL) (Richards & Schmidt, 2010), a proficiency test;

> measures how much of a language someone has learned. The difference between a proficiency test and an achievement test is that the latter is usually designed to measure how much a student has learned from a particular course or syllabus. A proficiency test is not linked to a particular course of instruction, but measures the learner's general level of language mastery. Although this may be a result of previous instruction and learning, these factors are not the focus of attention. (p. 464)

Proficiency tests are used to measure people's general language proficiency "prerequisite to entry or exit from some type of institution" (Brown, 1996, p. 9). Hughes (2003) points out that proficiency tests measure what people can do in the language; hence, their previous education and the content of language courses they have taken are not considered during assessment. In other words, while evaluating the general language ability of the test-taker, decisions are not based on specific syllabus. In proficiency tests, proficient means being proficient in the language for a specific purpose such as being proficient enough to follow a course in specific subjects like science, arts, or being good enough to do a study and follow a course at a university, or to work at an international corporation (Hughes, 2003). Some examples of proficiency tests used for these purposes are the internationally administered the Test of English as a Foreign Language (TOEFL) and the International English Language Testing System (IELTS). In Turkey, the Interuniversity Foreign Language Examination (ÜDS) and English proficiency exam for state employees (KPDS) are administered for the purposes mentioned above.

According to Brown (1996) proficiency tests are conducted “when a program must relate to the external world in some way” (p. 10). Other than the standardized tests such as TOEFL, IELTS, schools sometimes develop and conduct their own proficiency tests to decide (a) whether the students can fit into the program, and (b) whether they are proficient enough to succeed in other institutions with their existing language proficiency (Brown, 1996). While the former decision is made by conducting the proficiency test before entry, the latter decision is made based on the proficiency scores the students get from the test administered at exist.

As it is seen in Figure 1 below, these four types of tests are administered for different purposes. For this reason, extreme care must be exercised in developing, administering and scoring each test. For example, a proficiency test can be used to determine if the student is proficient enough to be accepted to a program; if he is not, a placement test should be administered to determine the proficiency level from which he/she should start the language course (Brown, 1996). However, while administering the proficiency tests and making decisions, learners’ background knowledge and previous training in that language should not be considered (Hughes, 2003) since they are designed to determine the general language ability of test-takers. Figure 1 presents the points to be considered before deciding to use any of the four language tests.

Type of Decision				
	Norm-Referenced		Criterion-Referenced	
Test Qualities	**Proficiency**	**Placement**	**Achievement**	**Diagnostic**
Detail of Information	Very General	General	Achievement	Diagnostic
Focus	Usually, general skills prerequisite to entry	Learning points all levels and skills of program	Terminal objectives of course or program	Terminal and enabling objectives of courses
Purpose of Decision	To compare individual overall with other groups/ individuals	To find each student's appropriate level	To determine the degree of learning for advancement or graduation	To inform students and teachers of objectives needing more work
Relationship To Program	Comparisons with other institutions	Comparisons within program	Directly related to objectives of program	Directly related to objectives still needing work
When Administered	Before entry and sometimes at exit	Beginning of program	End of courses	Beginning and/or middle of courses
Interpretation of Scores	Spread of scores	Spread of scores	Number and amount of objectives learned	Number and amount of objectives learned

Figure 1. Matching Tests to Decision Process. (Adapted from Brown, 1996, p. 9)

As highlighted by Bachman and Palmer (1996), if there is a mismatch between the test construct, the intended purpose of administering that specific test, and evaluation of assigned scores, the test-takers, the teachers and the institutions can be affected negatively. For example, the test takers may be misplaced at a class which is not appropriate for their language proficiency, can fail a course when they could pass, or may not be accepted into a program; the teachers can misinterpret the test scores and adopt a teaching approach inappropriate to their learner groups; the institutions can make wrong decisions in terms of curriculum and testing practices.

Qualities of Tests

Since tests are used to make important decisions about learners, teaching practices, and language programs, it is acknowledged that the most important point to be considered while developing and administering a test is the purpose of using that specific test (Bachman & Palmer, 1996; Brown, 1996; Hughes, 2003). According to Bachman and Palmer (1996), the usefulness of a test is the most important quality of a test, and they suggest a test usefulness model as “the essential basis for quality control throughout the entire test development process” (p. 17). This model consists of six test qualities: authenticity, interactiveness, washback and impact, practicality, construct validity, and reliability.

Authenticity is “defined as the relationship between test task characteristics, and the characteristics of tasks in the real world” (Fulcher & Davidson, 2007, p. 15). In other words, it is related to the extent to which the tasks are similar to the real-life situations. If a test requires the test takers to perform the tasks using real life language use (Bachman, 1990), it is considered to be authentic. Another quality of good tests is interactiveness which is defined by Fulcher and Davidson (2007) as “the degree to which the individual test taker’s characteristics (language ability, background knowledge and motivations) are engaged when taking a test” (p. 15). In other words, an interactive test requires the test-takers to use their individual characteristics to accomplish a test task. For example, a test task that requires a test-taker to activate his or her schemata, and relate the task topic to his or her existing topical knowledge is considered as an interactive task (Bachman & Palmer, 1996). A further quality of good tests, washback, also known as backwash, refers to the positive or negative effects of testing on teaching and learning (Hughes, 2003). Tests may also have impacts “on society and educational systems upon the individuals within those systems” (Bachman & Palmer, 1996, p. 29). Another quality of good tests is practicality which is different from the other five qualities in the sense that while those five qualities are concerned with the uses of test scores, practicality focuses on the development and administration of the test (Bachman & Palmer, 1996; Fulcher & Davidson, 2007). A test having the quality of practicality is easy and inexpensive to develop and administer. Validity, one of the most discussed qualities of tests, in general, refers to whether a test measures what it is supposed to measure (Brown, 1996; Hughes, 2003). It is also related to the extent to which interpretations of test scores are appropriate and meaningful. A test is said to be valid, if it assesses what it should assess. The term construct refers to a specific ability such as reading ability or listening ability for which a test task is designed to measure, and is used for interpreting scores obtained from this task. Therefore, the term construct validity is used to refer to the general notion of validity, and “the extent to which we can interpret a given test score as an indicator of the ability(ies), or construct(s), we want to measure” (Bachman & Palmer, 1996, p. 21). Last but not least, the final quality of good tests in Bachman and Palmer’s (1996) model is reliability which refers to “the consistency of test measurement” (p. 19). In other words, no matter when or where they take it, the test-takers will get the similar scores (Brown, 2004). Bachman (1990) suggests that instead of taking validity and reliability as two different aspects of measurement, they should be considered together in order to understand and control the factors that may affect test scores.

Reliability, for which Weir (2005) uses the term “scoring validity” (p. 22), is the focus of this study. Wiliam (2008) states that “A reliable test is one in which the scores that a student gets on different occasions, or with a slightly different set of questions on the test, or when someone else does the marking, does not change very much.” (p. 128) There are several types of reliability. For example, test-retest reliability is used for the consistency of test takers’ performance from occasion to occasion, and can be examined by giving the same test to the same group more than once (Bachman, 1990). Another form of reliability is the rater reliability which focuses on the consistency of raters especially when language tests are administered to assess written or spoken performance of test-takers and require human raters. The rater reliability is used for the raters' scoring performance, and can be measured in two ways. Inter-rater reliability, "the consistency of marking between markers" (Weir, 2005, p. 34), refers to the degree to which different raters agree on the scores they assigned. Intra-rater reliability, "each marker's consistency within himself" (Weir, 2005, p. 34), refers to the degree to which the same rater scores the same test similarly on two or more occasions. Hughes (2003) points out that when the decision or the result is very important for the test takers as it is in high stakes exams, achieving high reliability also becomes very important. However, it is not possible to entirely eliminate differences in assigned scores to a performance by the same rater or different raters (Bachman & Palmer, 1996). Yet, through careful test design and administration, the possible effects of the sources of inconsistency can be minimized.

Other than the six qualities suggested by Bachman and Palmer (1996), fairness in testing has also been considered as an important quality of good tests in relation to validity and reliability (Kunnan, 2000). According to Willingham and Cole (1997), fairness in testing can be assured by providing equal opportunities to candidates considering test design, test conduct, and scoring. The Code of Fair Testing Practices in Education (2004) prepared by the Joint Committee on Testing Practices is an important document that provides directions and standards for test developers and users related to the issue of fairness. The Code (2004) suggests that fairness should be considered in all aspects of testing process such as ensuring equal opportunities to every test-taker and reporting test results accurately.

Testing Speaking

Language assessment has gone through several changes, and recently performance assessment in which students are required to demonstrate the language skills they acquired has started to take the place of traditional test formats such as pencil-and-paper tests (McNamara, 1996). In this change, current trends in language teaching such as Communicative Language Teaching (CLT) have great impacts because teaching and testing cannot be considered as two separate things (Rudman, 1989). Speaking, one of the productive skills, has recently taken its place as an important part of curriculum in language teaching; thus, assessment of spoken language has also started to constitute an important component of English language assessment (Brown & Yule, 1999). In the last three decades, there has been a growing research interest in the development, implementation, and evaluation of tests which assess oral ability. In this section, formats of speaking assessment will be introduced, and factors that affect second language (L2) speaking skill assessment will be covered with an extensive focus on the rater effects on L2 speaking assessment.

Formats of Speaking Tests

Hughes (2003) remarks that there are three general formats of testing oral ability: “the interview, interaction with fellow candidates, and responses to audio-or-video-recorded stimuli” (p. 119). For assessing oral ability, Clark (as cited in O’Loughlin, 2001) has presented the distinguishing characteristics of three types of test format which are indirect tests, semi-direct tests, and direct tests. Similar to what Hughes (2003) suggests, these three formats have been widely acknowledged.

Indirect tests.

In indirect tests, test-takers do not need to speak and communicate. For this reason, these tests are considered as belonging to “pre-communicative era in language testing” (O’Loughlin, 2001, p. 4). Instead, the candidate, for example, can be asked to differentiate the pronunciation of different words. However, with recent trends in language teaching and testing which focus on the interaction and communicative skills, indirect tests of speaking where spoken language is not elicited are not preferable (Weir, 2005), and have been almost excluded from the language assessment practices.

Semi-direct tests.

In the semi-direct test format, language constructs can be elicited through the use of computer-generated or audio/video recorded stimuli to which the test-takers respond by using microphones (Hughes, 2003). Clark remarks that semi-direct tests are conducted in laboratories without a face-to-face communication and a live interlocutor (as cited in O’Loughlin, 2001). The tasks are presented thorough recordings, printed materials, and then, the candidate’s performance is recorded to be assessed by raters later. Due to the fact that the teaching of speaking skill has become necessary, so has the assessment of it. With the increasing importance given to speaking proficiency, McNamara (2000) suggests that the assessment of large numbers of candidates- feasibility can only be achieved through administering semi-direct tests. According to Hughes (2003), due to the necessity of testing many candidates at the same time, it can be economical if language laboratories are available. Moreover, with the growing interest in getting benefit from computer technology in delivering and administering tests (Qian, 2009), semi-direct tests have become a popular practice for professional testing organizations.

However, as Hughes (2003) asserts, semi-direct tests are inflexible in the sense that it is not possible to follow candidates’ responses because there is no interaction between the test-taker and the listener. These tests are less real life like due to the lack of interaction. In other words, they do not require the candidates to participate in a face-to face communication. Due their nature, (a) semi-direct speaking tests usually require the test-takers to speak in monologues; (b) there is no communicative and meaningful interaction between the candidates and other speakers; and (c) performing in such tasks can be more difficult than conversations for some language learners (O' Loughlin, 1997).

Direct Tests.

The direct tests or "live tests" (Qian, 2009, p. 114) were first used in the 1950s with the Oral Proficiency Interview (OPI) developed by the U.S. Foreign Services Institute (FSI), and since 1970s, OPI format has been widely used in the world to assess general speaking proficiency in a second language (O'Loughlin 2001). Direct tests are conducted as face-to-face, and test taker's performance is assessed by an interviewer. Thus, in literature, interview and direct-tests in oral assessment are used interchangeably. The interview, in which there is an interaction between the tester and the candidate, is the most commonly used format to test spoken proficiency of students (Hughes, 2003; Luoma, 2004). There are usually three participants in an oral interview: candidate is the test-taker; interlocutor or interviewer is the one interacting with the candidate; and examiner or rater is the person assessing the test-taker's performance (Alderson, Clapham, & Wall, 1995). In some cases, the interlocutor may also perform as a rater.

Direct-form of oral tests requires the test taker to perform oral tasks to demonstrate his or her oral language proficiency. Thus, it is also possible to take what Hughes (2003) suggests as a second type of speaking tests "interaction with fellow candidates" (p. 119) as a component of direct test format since there is a face-to-face interaction between the two candidates. According to Hughes (2003), the advantages of this format are as follows: the exchange of language utterances between the candidates is appropriate to their language level, and the candidates may perform better because they may feel more confident while speaking to an equal rather than to a superior, that is, the interviewer. However, in interviews, if students are expected to interact with a rater/ interlocutor or with a candidate with a higher proficiency level, it is possible that some of the language functions such as asking for information may not be elicited due to the fact that the candidates might feel like they are talking to a superior and may not be willing to take the initiative in the conversations (Hughes, 2003). Moreover, it is also possible that the performance of candidates can be affected from each other (a) negatively if paired with a personality wise dominant candidate who could dominate the discussion and do not let the other person take turns, and (b) positively if paired with a fellow candidate who can lead the discussion, guide and comfort his/her peer for better responses.

Several comparative studies have been conducted to investigate the advantages and disadvantages of direct and semi-direct tests (e.g., O'Loughlin, 2001; Oztekin, 2011; Qian, 2009). To assess oral proficiency, ideally, direct-tests in which candidates are assessed through spontaneous and face-to-face interaction (Lazaraton & Riggenbach, 1990) serve better to the notions of CLT. However, the practicality and feasibility of semi-direct tests can also make this format favorable especially for institutions with a large group of test-takers. For this reason, it is important for institutions to consider both the advantages and disadvantages of each format while choosing one or the other to assess spoken proficiency.

Moreover, there is also a growing interest in research to investigate the assessment and scoring procedures in oral interviews because of the discussions related to human factor. Human interaction in oral interviews is twofold; (a) candidate- interlocutor or rater interaction, and (b) candidate-candidate interaction. The next section will present the types of oral interviews in regards to the human interaction involved.

Types of oral interviews.

In terms of the number of test-takers they assess, oral interviews are conducted in three formats: individually, in pairs, and in groups. They are also grouped as oral interviews with single candidate, and oral interviews with multiple candidates.

In oral interviews where each candidate is assessed individually, the interaction takes place between the interlocutor and the candidate. In individual interviews, also referred to as one-to-one test, usually the interviewer starts the conversation and asks questions to find out the language proficiency of the candidate and to assess his or her performance.

Interview in pairs is another type of oral interviews during which the candidates perform a task which requires them to interact with each other (Luoma, 2004). In paired interviews, interlocutor observes the candidates rather than interacting with them directly. The task of the interlocutor is more difficult in this type of oral interviews because he or she has to make sure that each candidate understands the task, and pay attention to give equal time and opportunity for speaking to each candidate (Alderson et al., 1995). Similar to paired tasks, in oral interviews with group interaction task, there is candidate-candidate interaction, and the candidates are required to perform a group interaction task together.

Davis (2009) states that oral communication between peers takes place in many classroom and non-classroom speaking practices, “so use of pair work in assessment is well suited to educational context where the pedagogical focus is fully or partially task-based" (p. 368). This is also true for group interaction tasks in oral interviews since task based classroom activities are also practiced as group work. Hughes (2003) also suggests that, "if possible, it is desirable for candidates to interact with more than one tester" (pp. 124-125). Brooks' (2009) study investigating the effects of having a tester interlocutor (individual format) or another student (paired format) on test-taker's performance revealed that the students performed better in paired format than when they interacted with an examiner.

Several studies have revealed that candidates' performance may be affected negatively or positively from their interaction with other candidates. For example, the candidates' performance may be influenced by the other candidate's personality, communication style, and proficiency level (e.g., Davis, 2009; Iwashita, 1997; Luoma, 2004; O'Sullivan, 2002). Moreover, scoring procedures in oral interviews can be problematic. Assessing multiple candidates makes it more difficult to score each candidate's performance accurately and assign objective scores free from comparison of each candidate to his or her pair. Yet, factors that affect raters' oral assessment are not limited to the number of candidates.

Factors that Affect L2 Speaking Assessment

Several decisions are made based on students' language test scores (Brown, 1996). The purpose of language assessment studies is to "reduce sources of variability that are external to the learner's language performance to the greatest possible degree in order to reflect the candidate's true ability" (Wigglesworth, 2001, p. 188). With growing interest in CLT, performance assessment which requires human raters to assess the candidate's performance in a given writing or speaking task has become popular. However, studies revealed that, in performance assessment, there are some factors other than the candidates' performance that affect language test scores. There is a large body of research on writing assessment investigating the effects of some factors (e.g., the task, the scoring scale, the essay type) on candidates' performance and on raters' scorings (e.g., Carrell, 1995; Pula & Huot, 1993; Tedick, 1990; Weigle 1994, 1999). Since the focus of this study is related to the rater effects on oral interview scorings, the factors that will be discussed in the next paragraph come from research examining the factors that affect L2 speaking assessment. Before discussing these factors, it should be noted that because the assessment of L2 speaking performance has recently become necessary with the adaption of new approaches to language teaching, the theory and practice of testing L2 speaking proficiency, and the factors that affect L2 speaking performance assessment are still in the exploratory stage (Fulcher, 2003).

McNamara (1996) describes the interaction in performance testing and the affecting factors using a schematic representation (see Figure 2). The performance assessment in this model is composed of two processes: (a) the candidate's performing the task, and (b) the rater's assessing the performance. In an oral interview, the candidates with different backgrounds (candidate factor) perform a task (task factor) with or without an interlocutor/other candidates (interlocutor factor). In short, the performance process is affected by these three factors. Then, the rater (rater factor) scores the candidate's performance using a rubric (scale/criteria factor). Figure 2 presents the interaction of the affecting factors in performance testing.

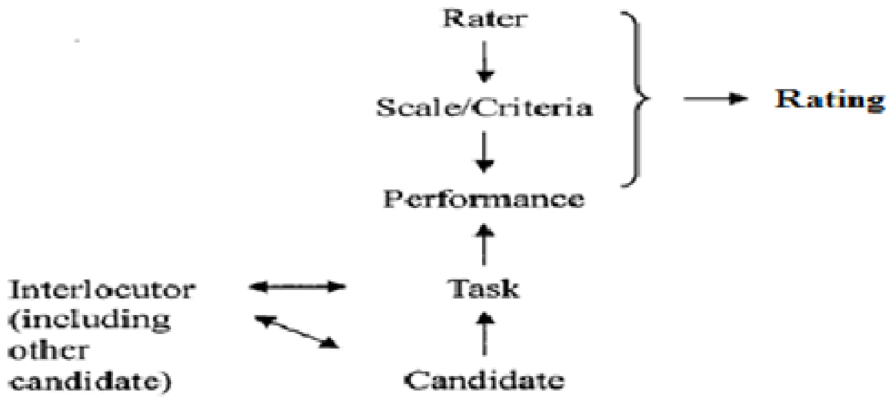

Figure 2. Proficiency and its relation to performance. (Adapted from McNamara, 1996, p. 86)

McNamara (1997) presents the notion of interaction in performance-based assessment by referring to several studies conducted on language testing. He states that the test takers are not the only affecting factors for the outcome of their performances; instead, interaction among other factors such as tasks, test formats, interlocutors, and raters should also be examined.

Bachman (1990) also suggests that there can be (a) test method factors such as the testing environment, the test rubric, (b) the examinees' personal attributes which are not related to their language ability such as cognitive style, sex, and ethnic background, and (c) random factors such as unpredictable testing conditions. He concludes that as the proficiency level of each candidate differs from one another, so do the effects of these factors on test performance of each candidate.

Brown (1996) emphasizes that the performances of test-takers on a given test can differ from each other, but their performances can also vary for several reasons. He groups these factors in two categories as "(1) those creating variance related to the purposes of the test (called meaningful variance here), and (2) those generating variance due to other extraneous sources (called measurement error, or error variance)” (Brown, 1996, p.186). Meaningful variance is about test validity, and defined as the variance that is directly related to the testing purposes. However, measurement error is “the variance in scores on a test that is not directly related to the purpose of the test” (Brown, 1996, p.188).

Brown (1996) divides measurement error into five categories according to the source of the error. The first source of measurement error, variance due to environment involves environmental factors such as noise, lighting, and weather that affect the students’ performance on a test. The second source, variance due to administration procedures, is related to the test administration procedures such as unclear or wrong directions for answering the questions and timing. For example, the studies comparing the administration of direct versus semi-direct methods of L2 speaking proficiency tests fall into this category (e.g., Stansfield & Kenyon, 1992). The effects of these two sources of measurement error are relatively controllable compared to the other three.

Variance attributable to the test and test items includes factors related to the test itself such as the clarity of the booklet, the format of the exam paper, and the number of items. Several studies have revealed the effect of tasks on test scores. For example, in oral proficiency assessment, task difficulty is the most often observed source of effect on test scores (Upshur & Turner, 1999).

Variance attributable to examinees, on the other hand, is about the condition of students such as their physical characteristics, psychological condition, and class or life experiences. According to Brown (1996), this variance constitutes a large part of the error variance.

O’Sullivan (2002) investigated the effects of test-takers’ familiarity with other candidates on their oral proficiency test pair-task performance. Thirty-two

Japanese university students with different proficiency levels performed in two pair work activities, one with a friend and one with a person they were not familiar with. The comparison of the candidates' performances in these two activities revealed that both exam partner's gender and proficiency level affect the pair-work language task performances of test-takers. The students who were acquainted with their partners scored better, and also when they worked with a partner with higher proficiency level, they performed better. As a result, O'Sullivan (2002) suggested that the acquaintanceship of the candidates should be considered not only while preparing and assessing any test that necessitates interaction between test-takers and/or interlocutors, but also during the pairing of the test-takers.

According to Brown (1996), the last source of measurement error, variance due to scoring procedures, is related to the factors that affect scoring procedures. For example, the use of holistic or analytic scales may affect scoring. As they are used to guide the raters while assigning scores, rating scales are significant in performance assessment. However, even when using the same rubric, raters may assign different weight to different components of the scale. In this case, the interpretation of scale components can cause measurement error. As a result of human errors in scoring, subjective scorings, variance in judgments, rater bias towards sex, race, age, and personality of the candidates, and rater characteristics such as severe rating tendency, the scoring of students' performances can be affected positively or negatively.

All the factors mentioned above are sources of measurement error that affect test-takers' scores. Due to the fact that testing of spoken language to assess communicative competence is open to raters' interpretations and rating differences (Ellis et al., 2002), concerns about validity, reliability, and fairness which are important qualities of a test (Bachman & Palmer, 1996; Kunnan, 2000) have been the center of the discussion about oral interviews for a long time (Joughin, 1998).

Raters

There are several studies conducted on the issue of subjective scoring as subjectivity in assessment contradicts with the qualities of a good test such as validity, reliability, and fairness. These studies have investigated the factors that affect the raters' scoring behaviors by referring to this phenomenon as rater variability, rater effect, measurement error, and rater bias (Myford & Wolfe, 2003).

In the findings of these studies, differences in rater scoring behaviors and assigned scores have been observed due to the existence of rater effects such as subjective scoring and rater bias. Bachman, Lynch, and Mason (1995) point out that due to its nature, in performance assessment, "potential variability in tasks and rater judgment, as sources of measurement error" can be observed (p. 239). More importantly, these studies of performance assessment of L2 proficiency have revealed that rater effects are systematic rather than random (Upshur & Turner, 1999). According to Crocker and Algina (1986), "sources of random errors include guessing, distractions in the testing situation, administration errors, content sampling, scoring errors, and fluctuations in the individual examinee's state" (p. 106). For example, tiredness can be a source of random error. On the other hand, if similarities are observed in rater's scoring behaviors, that is, if there is a pattern in relation to the measurement error, if the same type of error occurs consistently, then there is a systematic error rather than a random error. According to Haladyna and Downing (2004), "systematic error is not random, but group- or person specific" (p. 18). It is now acknowledged that some raters may show higher degree of severity in their judgments than other raters (Lumley & McNamara, 1995; Wigglesworth, 1993), but if there is a pattern in their behaviors towards a particular group of candidates, particular performance, particular task, this a crucial problem because this is a source of systematic measurement error. For example, if a rater consistently assigns lower scores to a certain group of candidates such as with the same race or gender, it is an act of systematic error in assessment, also referred to as rater bias (Linacre, 1994).

Types of rater effects on scores.

In performance assessment, rater variance “contributes to construct-irrelevant variance which can adversely affect an examinee’s test score” (Farrokhi & Esfandiari, 2011, p. 1532). In other words, the assigned scores may be the result of some systematic measurement errors which are not related to the assessed task. Studies have revealed that in performance assessment, raters, regardless of the training provided, seem to apply subjective scoring rather than applying the criteria (Brown, 1995; McNamara, 1990). For this reason, the variability in the assigned scores due to rater effects threatens validity, reliability and fairness of the scorings (Bachman, 2004; Eckes, 2005). There are four main types of rater effects: halo effects, central tendency, restriction of range, and severity/leniency (Saal, Downey, & Lahey, 1980).

Halo effect.

Of all the rater effects, halo effect is the most widely studied in the research literature (Myford & Wolfe, 2003). The term was coined by Thorndike in 1920 and defined as “a marked tendency to think of a person in general as either good or rather inferior and to color judgments of the qualities by their general feelings” (as cited in Farrokhi & Esfandiari, 2011, p. 1532). Borman describes it as “a tendency to attend a global impression of each examinee rather than to carefully distinguish among different levels of different performance dimensions” (as cited in Saal et al., 1980, p. 415). In other words, the assessment of one trait of the candidate can affect the assessment of his or her other traits. For example, a rater may be affected from the good vocabulary knowledge of a candidate for which he or she assigns a high score, and then may also assign a high score to the candidate's other traits such as grammar and pronunciation.

Central tendency.

Some raters may show evidence of central tendency which is "the rater's reluctance to make extreme judgments about other individuals" (Saal et al., 1980, p. 417). In other words, instead of using the lowest or highest scores in each category when necessary, the raters may overuse the middle categories of rating scales. Novice raters and raters who do not want to stand out usually yield to the effect of central tendency.

Restriction of range.

Restriction of range is similar to the central tendency effect in the sense that, regardless of candidate's performance, raters may tend to use certain scores in each category of the scoring rubric more often. While central tendency effect causes scores to cluster around midpoint, due to the restriction of range effect, raters assign scores usually around any particular point of the scale (Myford & Wolfe, 2003).

Severity / Leniency.

Rater severity or leniency is the rater tendency to assign scores from the lowest or highest bends of the scoring rubric categories. While severity is harsh rating, leniency is about being more tolerant and favorable during scoring. Researchers investigating rater effects focus more on rater leniency and severity because this is a very important factor in the inconsistency among raters, that is, inter-rater reliability. Raters may be severe in particular categories of the rubric such as grammar, pronunciation due to their perceptions about language teaching, or may show severity in all categories. If systematic, this type of rater effect has also shown evidence of rater bias towards a particular group of candidates. While Marr's study of the stability of rater severity revealed that rater severity is a random measurement error (as cited in Myford & Wolfe, 2000), Lumley and McNamara (1995) observed that the effects of a rater training session did not endure long, and the raters started to tend to score severely again after a while.

Factors that affect raters' scores.

Although the extent to which assessment scores are affected (e.g., halo effect, severity/leniency) is discussed quite in detail in the literature (e.g., Lumley & McNamara, 1995; Myford & Wolfe, 2000), the construct-irrelevant factors that affect raters' behaviors, scoring process, and final scorings in oral assessment have not been completely explored (Boulet, Van Zanten, McKinley, & Gary, 2001; Kang, 2012; Stoynoff , 2012). Since the focus of this study is to examine the effect of a particular factor on raters' scores, that is, the effect of the knowledge of candidates' proficiency level, some of the factors that affect raters' scores and are acknowledged to be significant by most of the researchers will be discussed in more detail below.

Raters' educational and professional experience.

Some studies investigating the rater effects on assigned scores focused on the effects of formal training in language (e.g., Brennan & Brennan, 1981; Chalhoub-Deville, 1995; Galloway, 1980; Thompson, 1991). Some of them also investigated the effects of teaching experience in ESL or EFL context (e.g., Chalhoub-Deville, 1995). Most of these studies suggested that listeners with language teaching experience were more severe in their ratings, especially about candidates' grammar (Hadden, 1991).

Thompson (1991) investigated the effects of raters' professional background on scores. Examining the scores assigned by language experts and inexperienced native speakers of English to the speech samples of 36 Russian candidates who speak English fluently, the researcher found out that experienced raters scored the accent category of the rubric more leniently than the raters without language-related training, and their reliability was higher. The researcher commented that the language training may increase raters' tolerance towards the foreign accent.

Chalhoub-Deville (1995) examined the behavior of three groups of raters who were 82 native speakers of Arabic from different professional backgrounds: 15 native speakers of Arabic teaching Arabic as a foreign language in the U.S., 31 non-teaching Arabs residing in the U.S. for at least one year, and 36 nonteaching Arabs living in Lebanon. Six subjects who were studying Arabic as a foreign language at a college were asked to participate in three tests: an oral interview, a narration and a read aloud. The results showed that the three rater groups paid attention to different aspects of language production although they used the same holistic rubric. Teacher raters tended to rate grammar more severely while non-teachers tended to focus on the more communicative aspects of the language performance.

Raters' nationality and native/ L2 language.

There have been several research investigating rater effects in oral assessment in relation to the L1 and L2 background of raters and candidates (e.g., Carey et al., 2010; Chalhoub-Deville & Wigglesworth, 2005; Derwing & Munro, 1997; Winke & Gass, 2012; Winke et al., 2011). Most studies investigating the differences in scores assigned by native speaker (NS) raters and nonnative speaker (NNS) raters have focused on NS and NNS teachers' ratings of NNS students' speech performances (e.g., Fayer & Krasinski, 1987; Hadden, 1991; Kim, 2009). The results of these studies have been inconclusive due to the fact that while some of these studies revealed that NNS raters tended to be more severe in their scorings than NS raters (e.g., Fayer & Krasinski, 1987), others suggested the opposite (e.g., Barnwell, 1989; Hill, 1996).

Currently, there is a growing interest of research on the effects of raters' familiarity and interaction with NNS of English with different L1 backgrounds such as Chinese, Japanese. In other words, Word English varieties have been a popular research interest (Gass & Varonis, 1984; Munro, Derwing, & Morton, 2006; Powers, Schedl, Wilson-Leung, & Butler, 1999). The findings are again contradictory in the sense that while some suggested that the familiarity affected listeners (e.g., Winke & Gass, 2012), some studies revealed no such effect (e.g., Munro et al., 2006).

Chalhoub-Deville and Wigglesworth (2005) investigated the effects of raters' nationality on their perceptions of speaking proficiency. The 124 raters from four English speaking countries, the U.S. (29), Australia (29), the UK (30), and Canada (35), were asked to assess TOEFL speaking tests of 12 international language students from six different language backgrounds. The researchers found that the UK raters were the most severe ones in their ratings while the U.S. raters were the most lenient ones.

Winke et al. (2011) investigated the effects of raters' familiarity with candidates L1 on the assigned scores. The TOEFL IBT (Internet-based test) speech samples of 72 test takers were rated by 107 raters who spoke Spanish, Korean, or Mandarin Chinese as L2. Using a many-facet Rasch measurement (MFRM) model, the researchers revealed that raters who speak Spanish as L2 assigned higher scores to the candidates whose L1 was Spanish than the candidates with the other two L1. The scores assigned by raters who speak Chinese as an L2 were also significantly higher for the candidates with Chinese L1 background. However, the raters who speak Korean as an L2 did not show significant leniency towards the test takers with Korean as an L1. The researchers also gathered qualitative data from the 26 of the raters by conducting stimulated recall sessions and observed that 15 raters referred to the accents of the test takers providing positive or negative remarks such as the accent was good or it made scoring difficult. The researchers concluded that raters' L2 background affected their judgments during assessing the candidates with a familiar language background.

Rater training.

The evidence of rater effects on ratings has led many institutions to provide rating training "to reduce both variability associated with differences in overall severity, and randomness" (Lumley & McNamara, 1995). In rating training sessions, first, the assessment criteria is introduced, and then the raters are asked to employ the criteria while rating some carefully pre-selected performances. The rating session is followed by a discussion of the assigned ratings to ensure consistency within the raters themselves and among different raters.

Most of the studies investigating the rater effects in relation to rater training (e.g., Lumley & McNamara, 1995; Myford & Wolfe, 2000) revealed that the training sessions were helpful in reducing random error in rater judgments and rater severity, and effective in ensuring raters' self-consistency; however, the positive effects of training sessions were not observed after a time-interval.

Lumley and McNamara (1995) examined the effect of rating training on the stability of rater severity and rater bias in three scoring sessions. Over a period of 20 months, they conducted two rater training sessions with an 18-month interval, and a subsequent test administration session. Four raters were asked to rate the performance of 11 test-takers with different health professions (e.g., doctors, nurses) who took the Occupational English Test to be allowed to practice in Australia. The findings of the study revealed that rater training to reduce rater severity was not effective in the long run, and that is why the researchers suggested that certified raters should be recruited for test administrations.

In fact, Lumley and McNamara's (1995) study is more related to rater characteristics and rater bias in terms of rater severity and the stability of rater severity. Some other studies focused on the effects of being experienced and inexperienced raters in relation to the notion of native speaker (e.g., Barnwell, 1989) which can be considered as a sub-category of the effects of raters' educational and professional experience and/or raters' nationality discussed before.

Candidates' and/or interviewers' gender.

There is a great deal of research on gender in second and foreign language education (Sunderland, 2000). In language assessment research, the studies have revealed that gender of the candidates and/or interviewers is another variable that may affect rating behaviors (e.g., Gholami, Sadeghi, & Nozad, 2011; O'Loughlin, 2002; O'Sullivan, 2000). According to Sunderland (2000), "Male and female interviewer's different styles and their different behavior towards male and female interviewees can be one possibility of gender effect in oral interviews" (as cited in Gholami et al., 2011, p. 1394), and this can affect the results of the interview. Most of the researchers have concluded that gender is one of the most significant variables that may affect raters' scorings whether positively or negatively.

For example, O'Sullivan (2000) looked into the impact of the gender of test-taker in relation to the interlocutor in oral proficiency interviews, and found out that female raters assigned higher scores to both male and female Japanese EFL learners than male interviewers, and they were more supportive by expanding their questions.

Other factors.

Since the factors that affect raters' scorings have not been completely explored yet, it is difficult to group all the studies in certain categories. In literature research, there are several studies that focused on a specific factor or several variables that affect rater behaviors. For example, Chuang (2011) investigated the effects of teachers' background differences on their ratings in oral proficiency assessments. The researcher focused on four specific rater-related variables: gender, native language, academic background, and training experience. The study revealed significant differences among raters in relation to those variables. For example, male teachers were harsher in their scorings. NS of English assigned lower scores than NNS raters. Teachers with no rating training on EFL speaking assessment tended to give higher scores than those who received training. The researcher commented that the most significant difference observed in the effects of major background related variable. Raters with linguistics or literature major backgrounds were more severe in their ratings than the raters with TESOL and other major backgrounds. Chuang (2011) concluded that these four background characteristics of raters were influential on test score differences.

The fact that rater effects can lead to misinterpretations and misjudgments of test-takers' actual performances, and thus, affect their academic success and future, has generated the need to further explore these construct-irrelevant factors in order to assure reliable, fair scores in high-stakes exams, and same applies for oral interviews in proficiency exams. Brown (1996) depicts the problem as follows: "The subjective nature of the scoring procedures can lead to evaluator inconsistencies or shifts having an effect on students' scores and affect the scorer reliability adversely" (p. 191). As discussed earlier, the most significant problem in testing speaking ability is reliability because raters may show evidence of variations in their judgments in assessing the performance of different candidates (Ur, 1999). In other words, in performance assessment, human raters may yield to subjectivity which is unwarranted in ensuring valid, reliable and fair test scores. For this reason, several measurement approaches have been adopted to investigate the effects of construct-irrelevant factors in performance assessment and to decide if the inconsistencies are systematic rather than random.

Existing Measurement Approaches to Test Rater Reliability

Researchers investigating rater effects use statistical analysis due to the fact that these systematic measurement errors can be observed by looking at the distribution of scores on a rating scale. SPSS and FACETS are the two most commonly used software to examine the correlations of assigned scores and detect rater effects. By entering the assigned scores into the software and considering other variables such as tasks, rubric, and raters, the studies have investigated rater variability from different perspectives such as rater severity/leniency.

Since performance assessment requires raters' judgment and interpretation of candidate's degree of success in performing the task and thus will "be subject to disagreement" (McNamara, 1996, p. 117), raters experience a complex procedural and cognitive process while assigning scores (Joe, 2008). In other words, human scoring involves two important principles "what raters perceive and how raters think" (Joe, 2008, p. 4). However, since the studies investigating rater affects have usually adopted quantitative data collection and analysis, the results have revealed that statistical approaches fail to understand what raters think during assigning scores, and why there are differences between two scores assigned to the same performance by the same rater (intra-rater reliability) and by different raters (inter-rater reliability). For this reason, qualitative analysis models such as observations and interviews have been adopted in recent studies. Recent studies have started to grow interest in applying especially cognitive processing models in order to gain better insights into how raters assign scores, and why there are differences among raters' scorings.

Verbal Report Protocols: Think-Alouds

Verbal report analysis, a frequently used method of exploring cognitive processes, has two types: concurrent verbal reports, also referred to as think alouds, are conducted simultaneously with the specified task, and retrospective verbal reports are gathered right after the performance task (Ericson & Simon, 1980). As it is sometimes difficult to remember what someone did and why he/she did it (Van Someren et al., 1994), think aloud protocols serve better in understanding raters' cognitive processing during oral assessment scoring. Ericson and Simon (1980) suggest that concurrent verbal protocols are more likely to provide reliable information because they report on an ongoing cognitive process (Kuusela & Paul, 2000), but retrospective verbal protocols rely on what raters can remember about what they thought during the task (Joe, Harmes, & Hickerson, 2011). As Joe (2008) and Joe et al., (2011) point out, the studies investigating rater cognition in oral assessment (e.g, Joe, 2008; Joe et al., 2011; Orr, 2002) have augmented our knowledge about rater cognition because to maintain fair assessment, it is important to understand how raters' prior knowledge and expectations affect their behaviors and scorings during oral interviews.

Orr (2002), for instance, investigated the decision making processes of 32 raters in the Cambridge First Certificate in English Speaking test. The raters were asked to provide verbal reports while assigning scores to four candidates' performances. The study revealed that the raters differed (a) in the severity of their judgments, (b) in the way they used the scoring criteria, and (c) in the way they referred to other factors which were not related to the assessment criteria. Orr's (2002) study, through the use of verbal reports, provided better insights into the complex decision-making and scoring process that all the raters go through in every assessment session.

As Fulcher (2003) proposes, in assessment of oral skills, it is not possible to assign ultimately reliable scores because the process is dependent on human raters who can be affected by several uncontrollable factors. However, the attempts to minimize the effects of these factors are noteworthy in order to have more valid, reliable and fair tests (Bachman, 1990; McNamara, 1996, 1997). In this respect, think aloud protocols serves better in providing great insights into what raters think during scoring process.

Conclusion

This chapter has reviewed the literature related to this study investigating the possible effects of the prior knowledge of raters of students' proficiency levels during oral interviews in proficiency exams. In this chapter, first, literature related to language testing has been reviewed by focusing on the four common types of tests and the qualities of tests. Next, the assessment of speaking ability and the formats of speaking tests have been discussed. Then, factors that affect speaking assessment have been elaborated by summarizing the relevant literature. Finally, existing measurement approaches to test reliability have been briefly introduced.

The next chapter will focus on the methodology of the study which covers the participants, setting, instruments, data collection procedures and data analysis.

CHAPTER III: METHODOLOGY

Introduction

The present study focuses on scorer reliability, particularly intra-rater reliability in oral interview assessments. The purpose of this study was to investigate the possible existence of rater bias and the effect(s), if any, of raters' prior knowledge of students' proficiency levels on their scorings.

In this respect, this study addresses the following research question:

- To what extent does raters' prior knowledge of students' proficiency level influence their assessment behaviors during oral interviews?

This chapter consists of five main sections: the setting and participants, the research design, instruments, procedure, and data analysis. In the first section, the setting and participants of this study are introduced and described in detail. In the second section, the research design that was employed in this study is explained briefly. In the third section, two different instruments, which are rating materials and data collection instruments are presented in reference to the research design. In the fourth section, the steps that are followed in the research procedure including the selection of participants and data collection are stated step by step. In the final section, the overall procedure for data analysis is provided.

Setting and Participants

The setting of this study is a preparatory school in a public university which provides intensive English courses to undergraduate students for one year. The students are required to take and pass the proficiency exam administered at the end of the academic year in order to pursue their studies in their departments. The rationale for choosing this school is both its providing convenience sampling to the researcher and its being one of the few public universities that administers oral interviews as a part of their proficiency exam and records and saves these oral interviews in their archives.

The participants of this study are 15 instructors who are native speakers of Turkish and teach English as a foreign language at the above mentioned university while also presuming the role as a rater in the oral interviews conducted at this university. Once the necessary permissions were received from the university, the teachers were contacted via e-mail, and they were presented the informed consent form on the norming session[6] which will be explained in detail in the procedure section (see Appendix 1 for the informed consent form). The participants were chosen on a voluntary basis, and they were regarded as the representative of all the instructors at this university since the total number of instructors working at this university is about 50. The demographic information of the participants was collected via a questionnaire designed by the researcher. It includes questions about the participants' educational background and experience in teaching and testing speaking (see Appendix 2). Table 1 presents the demographic information about the sample of this study.

[6] More information will be provided in the procedure section.

Table 1

Demographic Information of the Participants

Background Information	N (15)	%
Gender		
Female	10	66.66
Male	5	33.33
Undergraduate Major		
ELT	8	53.33
Other	7	46.66
Master's Degree		
No	10	66.66
Continuing	2	13.33
Completed	3	20
Doctoral Degree		
No	13	86.66
Continuing	2	13.33
Teaching Experience		
1-5	8	53.33
6-10	6	40
11+	1	6.66
Scoring Experience		
1-5	13	86. 66
6-10	2	13.33

Research Design

This study relies on a mixed-methods quasi-experimental research design which combines both quantitative and qualitative research during data collection and/or data analysis (Dörnyei, 2011). The study has been designed to collect the data in three sessions: (a) the norming session held to inform the participants about the study, collect demographic information, get their consent, and standardization for scoring, (b) the pre-test in which the raters were asked to assign scores without the knowledge of the students' proficiency levels, and (c) the post-test in which the information about students' proficiency levels was provided for the raters without making them aware of the actual purpose of the study. The raters were informed that the students' levels were written in the post-test grading sheet because some raters asked for that information in the pre-test. Moreover, both in the pre and post-test, think-aloud sessions were held during which the raters' verbal reports were gathered. Figure 3 shows the procedure followed to conduct the study.

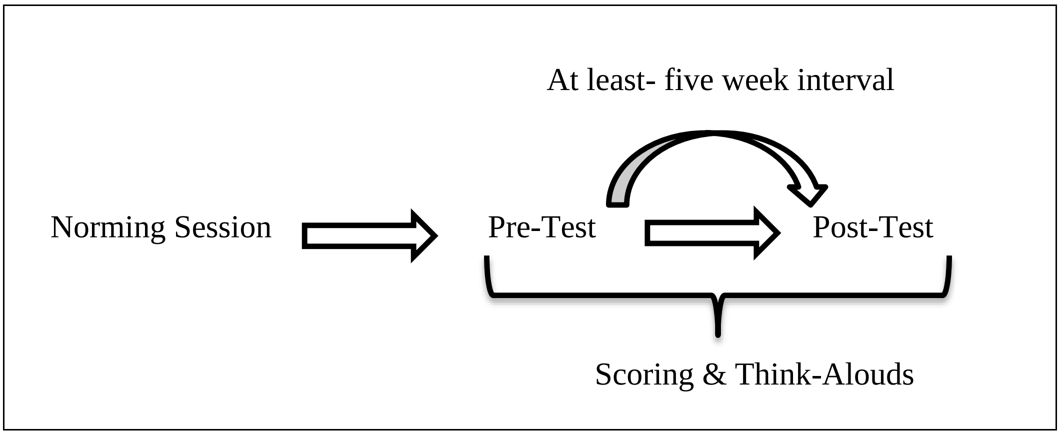

Figure 3. The procedure of the study.

While the scores assigned by the raters for each student's oral interview performance serve as the quantitative data, the raters' concurrent verbal reports provided during think aloud protocols for scoring process contribute to the study as qualitative data source in order to gain better insights about what raters think during scoring a performance. Thus, the qualitative and quantitative data are treated as complementary of each other during data collection and analysis.

Instruments

There are two kinds of instruments used for this study: (a) data collection instruments which are scores and think-alouds, and (b) rating materials, namely, video recordings, rating scale, and grading sheets (see Figure 4).

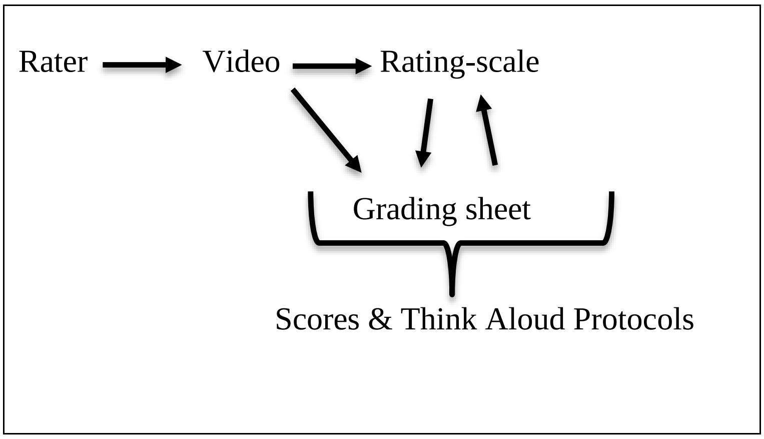

Figure 4. The interaction among the instruments during a scoring session conducted in this study.

As seen in Figure 4, data collection instruments which constitute the quantitative and qualitative data of this quasi-experimental research design are the records of raters' interactions with the rating materials.

Data Collection Instruments

Data collection instruments consist of two sets of data sources; (a) scores from the pre and post-test, and (b) concurrent verbal reports (think-alouds). These data indicate raters' evaluations of and judgments about students' spoken task performances in relation to the categories of the rating scale[7].

Scores.

The first set of data source, which are students' oral interview scores, are gathered during the pre and post-test which were conducted with at least five weeks interval. Using the rating materials, the raters individually assigned scores for each student twice, one for pre-test and one for post-test. The scores served as quantitative data for this study which were collected from raters under two conditions; first, raters' having no information about students' proficiency levels, and then, raters' being informed about students' proficiency levels both in written format and orally.

[7] More information about the rating scale will be provided in the rating materials section.

Concurrent verbal reports (Think-aloud protocols).

This set of data was complementary for the quantitative data, and gathered at the same time with the scores. They included approximately a five – minute - verbal reports of raters during which they commented on each student's performance while assigning scores right after watching the video recordings. Because both the researcher and the participants are native speakers of Turkish, the raters were asked to provide their verbal reports in Turkish so that they would feel more comfortable and provide more data. The raters' verbal reports were video-recorded, and in total, for pre-tests and post-tests nearly one-hour data was gathered from each rater, which added up to nearly 15 hours of recordings.

Rating Materials

Rating materials include those materials used by the raters while assigning scores. These materials consisted of (a) video recordings of oral interview performances of 12 students conducted as a part of 2011-2012 academic year proficiency exam, (b) the rating scale used by the raters while judging the student performance, and (c) the grading sheets for raters to fill while assigning scores.

Video recordings.

Six video recordings which were recorded during 2011-2012 academic year proficiency exam - oral interview sessions were chosen as a research instrument to let the raters assign scores for each student's oral interview performance. The video recordings were edited by the researcher in terms of deletion of the time allocated for students getting ready for the tasks. Thus, the length of each video were shortened to approximately seven minutes. Each video included oral interview session of two preparatory school students performing two tasks, one individually with the guidance of the interlocutor, and one interacting with the other candidate. In total, oral interview videos of 12 students with different proficiency levels[8] were used. There were four B level students, two C level students, and six D level students, and the students were randomly paired, either with a same proficiency level candidate, or with a higher or lower proficiency level one. Since these videos are kept in archives and no personal information about the students was given to the raters during this study, the students' consent for participation in the study was not taken. The same video recordings were used for the two scoring sessions which will be discussed in data collection instruments part below.

Rating scale.

In this study, the raters used the same analytic rubric used while assessing the oral performances of their students in the institution where the study was conducted. The analytic rubric, which was developed by the Speaking Office coordinator of the same institution, included five components which are *Fluency and Pronunciation, Vocabulary, Grammatical Range and Accuracy, Task Completion* and *Comprehension.* For each component, the lowest score that can be assigned is 1 point while the highest score is 4 points. As a *Total Score,* the raters can assign 5 points as the lowest score to a very poor performing student while the students with a successful performance can get up to 20 points (see Appendix 3).

Grading sheets.

Two grading sheets developed by the researcher were used by the raters while assigning scores in the pre and post-test. Although the same information about the students' pseudo IDs, the tasks they performed, and the categories of the rating scale was provided in the two forms, the proficiency levels of students which is an important feature of this research design were only presented in the grading sheet

[8] D/C/B levels: from the lowest to the highest proficiency levels.

used for the post-test. Moreover, in order to investigate whether the raters were familiar with any of the students, a section that asks whether the raters taught or knew the students was included in both sheets. The data gathered from the raters' scorings and verbal reports provided for the students with whom those raters were familiar were not included in the data analysis (see Appendix 4 and Appendix 5)

Data Collection Procedures

The researcher requested permission from the university to conduct the study, and after the permission has been received, first, an e-mail was sent to all the instructors working at the university the study was conducted in order to briefly inform them about the study and the procedure, and they were asked to respond to the e-mail pointing out whether they would like to participate or not. After the researcher received their responses, the instructors who accepted to participate were contacted face-to-face and invited to the norming session. Then, after asking the participants about a convenient time for them, the researcher scheduled the meeting.

All the participants attended the norming session. First, the informed consent form was given to the volunteers, and they provided their verbal and written consent to participate in the study (see Appendix 1 for the informed consent form).Then, the researcher made a PowerPoint presentation to the participants in order to give theoretical information about the study and to present the methodology of the study. The participants were informed about the amount of time necessary to be able to perform in the pre and post-test. They were not informed about the actual focus of the study which is the possible existence of rater bias and the effect(s), if any, of the raters' prior knowledge of students' proficiency levels on their scorings, but they were told that the researcher was interested in the process of raters' arriving at a decision for assigning scores.

Once they were informed about the study, for standardization, two pre-selected video recordings including two pairs of students' oral interview performances from 2011-2012 proficiency exam oral interview conducted at the institution were rated by the participants individually using the same analytic rubric chosen for this study. Since the raters were already familiar with using the rubric, no training was provided about the rubric, but the components and the descriptors the rubric includes were discussed very briefly. After the raters assigned scores for Video # 1, they were asked about what scores they assigned for each student in relation to the five components of the rubric and the *Total Score*. The scores were presented on the board in order to show the inconsistencies among the raters, and the reasons for the inconsistencies were discussed. The same procedure was followed for scoring Video # 2.

For pre-test scoring session, the participants were again asked about a convenient time for them during which they would perform individually as a rater. On the prescheduled day, the researcher conducted the pre-test with the participants whether at school in a quiet room, or at the house of the researcher due to time constraints. However, in both settings, the researcher paid immense attention to create the scoring atmosphere similar to actual oral interview assessments. Before the scoring procedure started for the pre-test Session 1, the raters were informed about think-aloud protocols, and they practiced scoring and providing verbal reports for one-preselected video which was not one of the six videos used as rating materials. After this practice session, the raters, first, watched one video, and then, provided verbal reports while scoring the students' performances. The same procedure was followed for each of six videos. The researcher was at present from the beginning to the end of the procedure as an observer; in other words, the researcher did not interfere with any part of the verbal reports unless there was a long pause while raters were talking about the students’ performances or the raters were likely to assign scores without verbalizing what they were thinking. If the raters’ did not provide verbal reports frequently without stopping, the researcher interfered by reminding the participant where he/she stopped using reflective phrases like *“You were saying….”* Moreover, the researcher tried to be as friendly as possible to make the participants comfortable while sharing their thoughts. No extra information was provided to the raters which might affect their judgments such as the scores assigned by the other raters, and the personal information about the students. Since in actual assessment, there is usually a break after five pairs of candidates, the raters were encouraged to take a five minute break after the fourth pair if necessary. The raters were not allowed to go back to the videos, rewind or forward it due to the fact that they are not able to go back to the speech samples of students during oral performance assessment. The order of the videos were assigned randomly for each rater in order to prevent future problems such as raters’ discussing about the videos with other participants although they were requested not to, and the order of the videos presented to the same rater were different in the pre and post-test in order to minimize the possible recall effect.

The same procedure with the pre-test was followed in the post-test scoring session which was conducted with at least a five-week interval. However, there was a major difference in terms of the information about the students provided to the raters. The difference is that the proficiency level of each student was written in the post-test grading sheet (see Appendix 5), and the raters were told that some raters asked for this information because, in actual assessments, they usually learn the students’ proficiency levels by looking at the exam documents. However, this was not the case. The information was purposefully kept in the pre-test and then provided in the post-test as a variable. The researcher made the explanation without overemphasizing it, and tried not to get participants' attention to the variable too much. Figure 5 shows the procedure followed during data collection.

Before pre-test **Norming Session**	**Pre-test**	**Post-test** **(at least 5 weeks Interval)**
• No training for rubric	• No information about students' proficiency levels	• Information about students' proficiency levels
• Two pre-selected videos for standardization	• Training for Think Alouds • Individual scoring for each student	• Reminding the procedure • Individual scoring for each student
• No information about the actual purpose of the study	• Each rater scores 12 students' performances	• Each rater scores 12 students' performances
• The raters were told that the focus was "the process of arriving at a decision for assigning scores"	• **Think Aloud Protocols** at the same time with scoring process (recorded)	• **Think Aloud Protocols** at the same time with scoring process (recorded)

Figure 5. Presentation of the research design in accordance with the procedure followed to collect data.

Data Analysis Procedures

Both quantitative and qualitative approaches to data analysis we used. The scores assigned by the raters were analyzed quantitatively including nonparametric statistics using the computer software in version 21 of SPSS while the video recordings of think aloud protocols were analyzed qualitatively.

First, the data collected via ratings were analyzed in version 21 of SPSS. The scores assigned by each rater were analyzed separately by using Wilcoxon Signed ranks test to see whether there is a significant difference between their pre and post test scores in the aspects of five categories of the rubric which are *Fluency and Pronunciation, Vocabulary, Grammatical Range and Accuracy, Task Completion, and Comprehension*, as well as in the *Total Scores*. Further analysis was also carried out with the rating data to investigate if the raters had bias towards students with a specific language proficiency level. The qualitative data gathered from think- alouds were analyzed with content analysis by using the framework of the rubric but also the other themes emerged which are not included in the rubric such as proficiency (see Appendix 6).

Conclusion

In this methodology chapter, the setting and participants, research design, instruments, and the procedure of the present study were described in detail. The next chapter will present detailed analysis of the quantitative and qualitative data gathered from the 15 participants through two complementary data collection instruments which are ratings and think-aloud protocols.

CHAPTER IV: DATA ANALYSIS

Introduction

The aim of the present study was to investigate the effects (if any) of raters' prior knowledge of students' proficiency levels on their scoring behaviors. In this respect, this study addressed the following research question:

- To what extent does raters' prior knowledge of students' proficiency level influence their assessment behaviors during oral interviews?

In this quasi-experimental study, 15 raters who have been teaching English as a foreign language (EFL) at a state university assessed the oral performances of 12 students twice by watching the same six pre-recorded videos, each of which included 2011-2012 proficiency exam oral interviews of two students. There was a five-week interval between the pre and post-test to avoid any recall effect. Two sets of data were collected in the pre and post-test for this study: (a) quantitative data consisted of the scores assigned twice by 15 raters to each student's oral interview performance, and (b) qualitative data gathered from the verbal reports of 15 raters in the pre and post-test think-aloud protocols while they were assigning scores. In accordance with the adopted mixed-methods research design, the data from the pre and post-test scores were analyzed quantitatively while the data from think-aloud protocols were evaluated qualitatively. This chapter will first introduce the data analysis procedures, and then the overall results of the quantitative data analysis will be presented. In the next section, quantitative data followed by the qualitative data from the raters' verbal reports will be discussed separately for the two groups of raters: (a) raters' who showed statistically significant differences in the scores they assigned, and (b) raters' who did not show significant differences in their scorings.

Data Analysis Procedures

After the pre-test and post-tests were administered, the quantitative data obtained from the scores assigned by 15 raters according to the criteria provided were entered into Statistical Package for Social Sciences (SPSS, version 21). Wilcoxon Signed Ranks Test, which is a nonparametric test for small sample sizes, was run for each rater's assigned scores in the pre and post-test in order to determine if there was a statistically significant difference between the scores assigned without the knowledge of students' proficiency levels (pre-test) and with that knowledge (post-test). The scores of the students with whom the raters were familiar with were not included in the data analysis in order to prevent the effect of familiarity with the students on the test results.

The rubric used in this study included five components which are *Fluency and Pronunciation, Vocabulary, Grammatical Range and Accuracy, Task Completion* and *Comprehension.* For each component, the lowest score that can be assigned is 1 point while the highest score is 4 points. As a *Total Score*, the raters can assign 5 points as the lowest score to a very poor performing student while the students with a successful performance can get up to 20 points.

After running the non-parametric Wilcoxon Signed Ranks Test for the pre and post-test scores assigned for each component of the rubric by each rater separately, the verbal reports provided by the raters during the pre and post-test scoring procedures were analyzed qualitatively by adopting the framework of the rubric used by the raters while assigning scores. Moreover, the themes emerged other than these five components of the rubric were also taken into consideration while doing the analysis.

Results

The results will be presented in accordance with the research question of the study, that is, "*To what extent does raters' prior knowledge of students' proficiency level influence their assessment behaviors during oral interviews?".* It should be noted that there were three levels in the institution where the study was conducted: D level is the lowest, C level is lower and B level is the highest proficiency level. The answer to the research question will be discussed in two sections. First, quantitative data gathered from each rater's assigned scores in the pre and post-test will be introduced to present those raters who had statistically significant difference in their scorings and who showed no significant difference in their scoring. The second section will focus on the analysis of the raters' verbal reports in relation to the assigned scores. In this section, the data will be presented in two parts. First, the data gathered from the raters who had statistically significant difference between their pre and post-test scorings will be discussed. There will be separate discussion for each of these raters. Then, the data from the raters with no significant difference between their pre and post-test scorings will be shown. In each part, the data from the raters who referred to the proficiency levels of the students in their verbal reports will be followed by the data from the raters who did not mention the proficiency levels of the students.

Pre and Post-test Data Analysis for 15 Raters

A Wilcoxon matched pairs signed rank test was conducted to determine whether there was a difference in the pre-test and post-test scores assigned by the raters (see Table 2).

Table 2

The Difference between Each Rater's Pre and Post-Test Ratings through Wilcoxon Signed Ranks Test

Rater	Z Asymp. Sig. (2-tailed)	Fluency and Pronunciation	Vocabulary	Grammatical Range and Accuracy	Task Completion	Comprehension	Total Score
1	**Z**	**-2.236[c]**	**-2.714[c]**	**-.378[c]**	**.000[b]**	**-2.070[c]**	**-2.552[c]**
	Asymp. Sig. (2-tailed)	**.025***	**.007***	**.705**	**1.000**	**.038***	**.011***
2	**Z**	**-.557[a]**	**-.541[a]**	**-1.552[a]**	**-.707[a]**	**-1.098[c]**	**-.835[a]**
	Asymp. Sig. (2-tailed)	**.577**	**.589**	**.121**	**.480**	**.272**	**.404**
3	**Z**	**-.816[a]**	**-.816[a]**	**-2.000[c]**	**-.816[a]**	**-.276[c]**	**-.155[c]**
	Asymp. Sig. (2-tailed)	**.414**	**.414**	**.046***	**.414**	**.783**	**.877**
4	**Z**	**-2.236[a]**	**-2.121[a]**	**-.378[a]**	**-.447[a]**	**.000[c]**	**-1.620[a]**
	Asymp. Sig. (2-tailed)	**.025***	**.034***	**.705**	**.655**	**1.000**	**.105**
5	**Z**	**.000[b]**	**-1.342[c]**	**-.447[a]**	**-.378[c]**	**-.447[c]**	**-.647[c]**
	Asymp. Sig. (2-tailed)	**1.000**	**.180**	**.655**	**.705**	**.655**	**.518**
6	**Z**	**.000[b]**	**-.577[c]**	**-.816[a]**	**-2.333[c]**	**-1.633[c]**	**-1.160[c]**
	Asymp. Sig. (2-tailed)	**1.000**	**.564**	**.414**	**.020***	**.102**	**.246**
7	**Z**	**.000[b]**	**-.632[a]**	**-1.000[c]**	**-.447[c]**	**-.722[a]**	**-.254[a]**
	Asymp. Sig. (2-tailed)	**1.000**	**.527**	**.317**	**.655**	**.470**	**.799**
8	**Z**	**-.447[a]**	**-.577[c]**	**-2.236[a]**	**-1.000[c]**	**-2.000[a]**	**-1.474[a]**
	Asymp. Sig. (2-tailed)	**.655**	**.564**	**.025***	**.317**	**.046***	**.140**

Rater	Z Asymp. Sig. (2-tailed)	Fluency and Pronunciation	Vocabulary	Grammatical Range and Accuracy	Task Completion	Comprehension	Total Score
9	Z	-1.342[a]	-.447[c]	-1.000[a]	-1.265[a]	-.707[a]	-1.012[a]
	Asymp. Sig. (2-tailed)	.180	.655	.317	.206	.480	.311
10	Z	-.577[a]	-1.414[a]	-.577[c]	-1.633[c]	-.447[a]	-.181[c]
	Asymp. Sig. (2-tailed)	.564	.157	.564	.102	.655	.856
11	Z	-1.134[a]	-.447[a]	-2.236[a]	-.707[a]	-.447[c]	-1.544[a]
	Asymp. Sig. (2-tailed)	.257	.655	.025*	.480	.655	.123
12	Z	-1.890[a]	.000[b]	.000[b]	-1.000[a]	-1.000[c]	-.568[a]
	Asymp. Sig. (2-tailed)	.059	1.000	1.000	.317	.317	.570
13	Z	.000[b]	.000[b]	-.966[c]	-1.000[a]	-1.406[c]	-.565[c]
	Asymp. Sig. (2-tailed)	1.000	1.000	.334	.317	.160	.572
14	Z	-1.414[a]	-2.236[a]	-1.732[a]	-.577[c]	-.577[c]	-1.930[a]
	Asymp. Sig. (2-tailed)	.157	.025*	.083	.564	.564	.054
15	Z	-.816[c]	-.378[c]	-.378[c]	-2.646[c]	-.378[c]	-1.702[c]
	Asymp. Sig. (2-tailed)	.414	.705	.705	.008*	.705	.089

a.Based on positive ranks

b.The sum of negative ranks equals the sum of positive ranks.

c. Based on negative ranks

*p<.05

As Table 2 shows, while negative ranks demonstrate that there was a decrease in the assigned scores, positive ranks show that the raters assigned higher scores in the post-test. Eight raters, Rater # 1, Rater # 3, Rater # 4, Rater # 6, Rater # 8, Rater # 11, Rater # 14, and Rater # 15, behaved differently while assigning scores in the post-test, that is the knowledge of the students' proficiency levels affected the scores they assigned in at least one component of the rubric. As far as the *Total Scores* are concerned, there was a significant difference between the pre and post-test only in the scorings of Rater # 1. The raters who showed consistency in their scoring behaviors in the pre and post-test are Rater # 2, Rater # 5, Rater # 7, Rater # 9, Rater # 10, Rater # 12, and Rater # 13.In other words, the knowledge of students' proficiency levels did not affect seven raters' scorings significantly. However, although the results of Wilcoxon matched pairs signed rank test indicated that there was a significant difference only in one rater's pre and post-test *Total Scores*, when the descriptives of the pre and post-test *Total Scores* assigned to individual students were analyzed, it was observed that the majority of the scores assigned by the 15 raters changed in the post-test as higher or lower *Total Scores* (see Table 3 and Figure 6).

Table 3

Comparison between the Pre and Post-test for the Total Scores

Raters	**Negative Ranks***	**Positive Ranks****	**Ties*****	**Scorings Included[9]**	**Scorings Excluded[9]**
1	**0**	**8**	**1**	**9**	**3**
2	**6**	**4**	**2**	**12**	**0**
3	**5**	**5**	**2**	**12**	**0**
4	**6**	**2**	**2**	**10**	**2**
5	**2**	**3**	**5**	**10**	**2**
6	**2**	**4**	**5**	**11**	**1**
7	**4**	**3**	**4**	**11**	**1**
8	**6**	**1**	**4**	**11**	**1**
9	**6**	**3**	**3**	**12**	**0**
10	**4**	**5**	**3**	**12**	**0**
11	**8**	**2**	**2**	**12**	**0**
12	**6**	**4**	**2**	**12**	**0**
13	**5**	**5**	**1**	**11**	**1**
14	**6**	**1**	**4**	**11**	**1**
15	**2**	**8**	**2**	**12**	**0**
TOTAL	**68**	**58**	**42**	**168**	**12**

* post-test scores < pre-test scores
** post-test scores > pre-test scores
*** post-test scores = pre-test scores

[9] The scores of the students with whom the raters were familiar with were not included in the quantitative data analysis in order to prevent the effect of familiarity with the students on the test results.

In both pre and post-test, a total of 180 scorings was done by 15 raters. In total, 168 scores were included in the analysis due to the fact that some raters reported that they were familiar with one or more of the students whose performance was assessed. In total, 12 scores assigned by these raters were excluded from the data for this reason. Table 3 shows the differences between the pre and post-test *Total Scores* assigned by each rater. Although 42 *Total Scores* did not change in the post-test, as shown in Table 3, while 58 scores increased, 68 scores decreased. Figure 6 presents the percentages of the negative ranks, positive ranks, and ties.

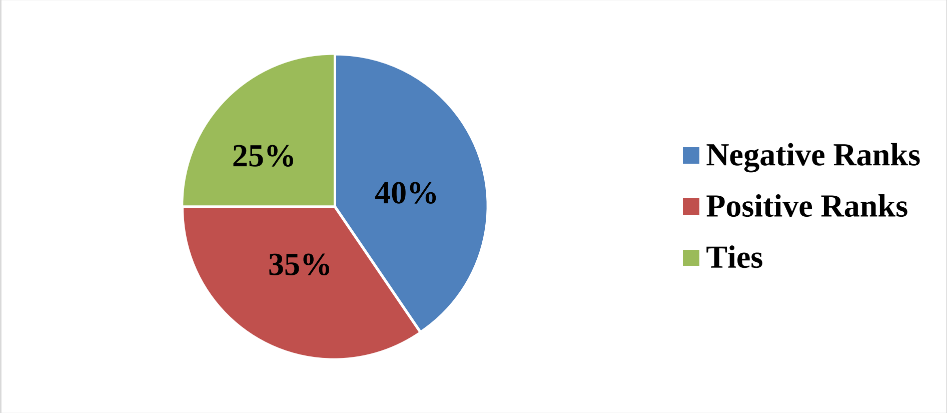

Figure 6. The percentages of negative ranks, positive ranks, and ties in the post-test *Total Scores*.

As seen in Figure 6, 25 % of *Total Scores* did not change while changes were observed in the 75 % of the assigned scores. While 40% of the *Total Scores* ranked lower, 35 % of the post-test *Total Scores* ranked higher than the pre-test *Total Scores*. When the scorings were analyzed separately for the raters with statistically significant differences and the raters with no significant difference in their scorings, the results indicated that there are similarities in terms of the percentages of the scores that did not change in the post-test; however, differences were observed for the negative and positive ranks (see Table 4 and Table 5).

Table 4

Comparison between the Pre and Post-test for the Total Scores Assigned by the Raters[a] with Significant Difference

	Negative Ranks *	**Positive Ranks****	**Ties*****	**Scorings Included[10]**	**Scorings Excluded[10]**
TOTAL	**35**	**31**	**22**	**88**	**8**
PERCENTAGES	**40**	**35**	**25**	**88**	**8**

* post-test scores < pre-test scores
** post-test scores > pre-test scores
*** post-test scores = pre-test scores
a. Raters # 1, # 3, # 4, # 6, # 8, # 11, # 14, # 15

Table 5

Comparison between the Pre and Post-test for the Total Scores Assigned by the Raters[b] without Significant Difference

	Negative Ranks*	**Positive Ranks****	**Ties*****	**Scorings Included[10]**	**Scorings Excluded[10]**
TOTAL	**33**	**27**	**20**	**80**	**4**
PERCENTAGES	**41**	**34**	**25**	**80**	**4**

* post-test scores < pre-test scores
** post-test scores > pre-test scores
*** post-test scores = pre-test scores
b. Raters # 2, # 5, # 7, # 9, # 10, # 12, # 13

As seen in Table 4 and Table 5, although the two groups of raters were more severe in their ratings in the post-test, the seven raters who had no significant difference in their scorings were slightly more severe than the eight raters who had significant differences in their scorings. Given these higher frequencies in negative and positive

[10] The scores of the students with whom the raters were familiar with were not included in the data analysis in order to prevent the effect of familiarity with the students on the test results.

ranks, for more in depth analysis, the verbal reports of the raters in relation to the scores they assigned will be analyzed in the next section.

Analysis of the Raters' Verbal Reports in Relation to the Assigned Scores

In this section, first, the data gathered from raters who showed significant difference between their pre and post-test scorings will be presented. Then, the data for raters with no significant scoring difference will be introduced.

Data analysis for the raters with statistically significant difference between their scorings.

The results indicated that eight raters, Rater # 1, Rater # 3, Rater # 4, Rater # 6, Rater # 8, Rater # 11, Rater # 14, and Rater # 15, did not show consistent scoring behaviors within themselves in different sections of the rubric (see Table 6).

Table 6

The Components of the Rubric in Which There Was a Statistically Significant Difference between the Pre and Post-test Scores Assigned by Each Rater

Rater	**Fluency & Pronunciation**	**Vocabulary**	**Grammatical Range & Accuracy**	**Task Completion**	**Comprehension**	**Total Score**
1	**.025***	**.007***			**.038***	**.011***
3			**.046***			
4	**.025***	**.034***				
6				**.020***		
8			**.025***		**.046***	
11			**.025***			
14		**.025***				
15				**.008***		

*p<.05

As shown in Table 6, each rater behaved differently while assigning scores to different components of the rubric. While some raters assigned different scores only in one feature, some raters assigned higher or lower scores in more than one component. The *Vocabulary* and *Grammatical Range and Accuracy* were the components of the rubric in which the raters showed significant differences the most frequently while one rater (Rater # 1) behaved differently in the *Total Scores* component. In this section, the data gathered from each rater's scorings and verbal reports will be presented separately. First, the data from the six raters who referred to the proficiency levels of the students will be introduced. Then, the data from the two raters who did not refer to the levels of the students in their verbal reports will be shown.

Raters who referred to the proficiency levels of the students.

Data analysis for Rater # 1.

The results indicated a statistically significant difference between the pre and post-test scores assigned by Rater # 1[11] in the *Fluency and Pronunciation* (z = -2.236, p = .025), *Vocabulary* (z = -2.714, p = .007), *Comprehension* (z = -2.070, p = .038) *and Total Scores* (z = -2.552, p = .011). However, the results of further analysis on the assigned scores revealed that there were also higher and/or lower rankings in the other components of the rubric (see Table 7).

[11] Rater # 1 reported that she was familiar with three students, Student # 1, Student # 8 and Student # 9, so the scores assigned to those students' performances were not included in data analysis.

Table 7

Comparison between the Pre and Post-test for all the Scores Assigned by Rater # 1[12]

Component	**Negative Ranks***	**Positive Ranks****	**Ties*****
Fluency and Pronunciation	**0**	**5**	**4**
Vocabulary	**0**	**8**	**1**
Grammatical Range and Accuracy	**2**	**2**	**5**
Task Completion	**2**	**2**	**5**
Comprehension	**0**	**5**	**4**
Total Score	**0**	**8**	**1**
TOTAL	**4**	**30**	**20**

* post-test scores < pre-test scores
** post-test scores > pre-test scores
*** post-test scores = pre-test scores

As Table 7 shows, Rater # 1 mostly assigned higher points in the post-test. She also assigned equal points for a great majority of the scorings while only four scorings ranked lower. Given these significant differences and the higher rankings, how different proficiency levels received different attention from Rater # 1 was analyzed (see Table 8).

[12] Rater # 1 reported that she was familiar with three students, Student # 1, Student # 8 and Student # 9, so the scores assigned to those students' performances were not included in data analysis.

Table 8

The Frequency of the Ranks in the Total Scores Assigned by Rater # 1[13] *for Each Level*

Levels[14]	**Negative Ranks ***	**Positive Ranks****	**Ties*****
D Levels	**0**	**2**	**1**
C Levels	**0**	**2**	**0**
B Levels	**0**	**4**	**0**
TOTAL	**0**	**8**	**1**

* post-test scores < pre-test scores
** post-test scores > pre-test scores
*** post-test scores = pre-test scores

As seen in Table 8, Rater # 1 was more lenient in the post-test scoring session when the information about the students' proficiency levels was available. The post-test *Total Scores* assigned by Rater # 1 for each level ranked higher than the pre-test scores except for only one student's: a D level student's post-test *Total Scores* did not change. When the verbal reports of Rater # 1 were analyzed to see whether she made references to students' proficiency levels in the components especially where significant differences were observed, it was found that she made several references to the students' proficiency levels. Figure 7 shows some extracts from the pre and post-test verbal reports of Rater # 1 in relation to the scores she assigned.

[13] Rater # 1 reported that she was familiar with three students, Student # 1, Student # 8 and Student # 9, so the scores assigned to those students' performances were not included in data analysis.

[14] D/C/B Levels (D: the lowest, B: the highest)

Student No	Student Level [15]	Partner's No & Level [16]	Component of the rubric	Pre-test score & comment	Post-test score & comment	Pre-test *Total Score*	Post-test *Total Score*
2	C	1 - D	*Vocabulary*	(2) *The student used very limited vocabulary*	(3) *It is clear that the student is a C level student and her vocabulary use/range was not bad*	(12)	(15)
3	B	4 - B	*Vocabulary*	(3) *The student's vocabulary use and range is not as good as her partner's*	(4) *It is evident that the student is a B level student, and she used accurate and appropriate words*	(17)	(18)
10	B	11 - D	*Total Score*	(5) *Very bad, no effort to speak*	(13) *Although B level, she could not speak, and could not do the task.*	(5)	(13)
			Grammatical Range and accuracy	(1) *There were almost no sentences at all.*	(3) *No big mistakes*		

Figure 7. Examples of assigned scores and verbal reports by Rater # 1.

No conclusion about how Rater # 1 assessed the students with different proficiency levels can be drawn from the data available. However, the

[15] D/C/B Levels (D: the lowest, B: the highest)

[16] This information is provided since some raters referred to the level of the students' partners and assigned scores by comparing them.

results indicated that a great majority of the scores assigned by Rater # 1 were favorable rankings in the post-test when the information about the students' proficiency levels were available to her.

Data analysis for Rater # 3.

The results indicated a statistically significant difference between pre and post-test scores assigned by Rater # 3[17] in the *Grammatical Range and Accuracy (*z = -2.000, p = .046) scores. While there was no change in the pre and post-test scores of eight students in this component, the rater assigned higher scores to four lower level students, three of whom were D level students and one was a C level student. However, when all the scores assigned by Rater # 3 were analyzed, it was observed that there were also higher and lower rankings in the other components of the rubric (see Table 9).

Table 9

Comparison between the Pre and Post-test for all the Scores Assigned by Rater # 3

Component	**Negative Ranks***	**Positive Ranks****	**Ties*****
Fluency and Pronunciation	**4**	**2**	**6**
Vocabulary	**4**	**2**	**6**
Grammatical Range and Accuracy	**0**	**4**	**8**
Task Completion	**4**	**2**	**6**
Comprehension	**2**	**3**	**7**
Total Score	**5**	**5**	**2**
TOTAL	**19**	**18**	**35**

* post-test scores < pre-test scores
** post-test scores > pre-test scores
*** post-test scores = pre-test scores

[17] Rater # 3 reported that she was not familiar with any of the students, so there is no missing data in this analysis.

As shown in Table 9, the rater assigned the same scores in nearly half of the scoring sessions. However, the total of negative and positive ranks were higher than the ties when all the scores were considered. Given these different rankings between her pre and post-test scorings, how different the *Total Score*s assigned for different proficiency levels were ranked by Rater # 3 was analyzed (see Table 10).

Table 10

The Frequency of the Ranks in the Total Scores Assigned by Rater # 3[18] *for Each Level*

Levels[19]	**Negative Ranks ***	**Positive Ranks****	**Ties*****
D Levels	**2**	**2**	**1**
C Levels	**0**	**2**	**1**
B Levels	**3**	**1**	**0**
TOTAL	**5**	**5**	**2**

* post-test scores < pre-test scores
** post-test scores > pre-test scores
*** post-test scores = pre-test scores

As seen in Table 10, Rater # 3 assigned lower and/or higher scores for 10 students while she did not change her scorings for two students. She assigned equally lower and higher scores for D level students. While she was more lenient while assessing the performances of C level students, she was more severe while she was assigning scores for B level students. Given these higher and positive ranks in the *Total Scores* of 12 students, the verbal repots of Rater # 3 were analyzed to see whether she referred to the proficiency levels of the students while she was assigning scores, especially in the *Grammatical Range and Accuracy* scorings in which a statistically significant difference was observed. Figure 8 shows some extracts form the pre and post-test verbal reports of Rater # 3 in relation to the scores she assigned.

[18] Rater # 3 reported that she was not familiar with any of the students, so there is no missing data in this analysis.

[19] D/C/B Levels (D: the lowest, B: the highest)

Student No	Student Level [20]	Partner's No & Level[21]	Component of the rubric	Pre-test score & comment	Post-test score & comment	Pre-test *Total Score*	Post-test *Total Score*
1	D	2- C	*Grammatical Range and Accuracy*	(1) *The student had several mistakes even in basic structures and simple sentences.*	(2) *The student had frequent mistakes and could not deliver the message. When we consider her level, she is a fair student in her level especially in the second task.*	(8)	(10)
			Comprehension	(1) *She had difficulty in understanding the both tasks. She had major problems in comprehension.*	(3) *She understood the message, but she needs repetition.*		
2	C	1- D	*Grammatical Range and Accuracy*	(3) *There were problems in the first task, but the second task was better although there were some mistakes.*	(3) *She had frequent errors, but it was not difficult to understand her sentences. When we consider her level, she is a poor C level student.*	(12)	(15)
4	B	3 - B	*Fluency and Pronunciation*	(4) *The student had some problems with the pronunciation of some words, but she was good, she had no hesitations in terms of fluency.*	(3) *She had some hesitations, wrong pronunciation for some vocabulary, but in general, she could deliver the message if we consider they are B level students.*	(20)	(18)

[20] D/C/B Levels (D: the lowest, B: the highest)

[21] This information is provided since some raters referred to the level of the students' partners and assigned scores by comparing them.

			Vocabulary	(4) *Both students were very successful.*	(3) *No problem. She used appropriate vocabulary. If we compare these two students, her partner used more conversational expressions.*		
7	D	8 - C	*Grammatical Range and Accuracy*	(2) *She had some errors/mistakes that obscured the meaning. Her partner was better than her in terms of grammatical range and accuracy.*	(3) *She had some problems, but they did not obscure the meaning. Her partner, the C level student, was better.*	(14)	(13)
8	C	7 - D	*Grammatical Range and Accuracy*	(3) *He was better than his partner in terms of word order in the second task.*	(3) *C level student was better. He had some mistakes, but they did not obscure meaning.*	(15)	(15)

Figure 8. Examples of assigned scores and verbal reports by Rater # 3.

The results indicated that out of 12 students, 10 students' *Total Scores* changed in the post-test. As seen in Figure 8, the rater referred to the students' levels and their partners' while assigning scores. The most significant difference was observed in the *Grammatical Range and Accuracy* scores. When the verbal reports for Student # 1 in relation to Student # 2, and Student # 7 in relation to Student # 8 were examined, it was found out that D level - that is the lowest level - students were assessed more favorably in the post-test in terms of *Grammatical Range and Accuracy* when the rater knew their levels and the fact that they were paired with a different level student. B level - that is the highest level - students were assessed more severely by Rater # 3 in terms of *Fluency and Pronunciation* (e.g., Student # 4 in Figure 8) in the post-test.

Data analysis for Rater # 4.

The results indicated a statistically significant difference between the pre and post-test scores assigned by Rater # 4[22] to *Fluency and Pronunciation* (z = -2.236, p = .025) and *Vocabulary (*z = -2.121, p = .034). When all the scores assigned by Rater # 4 were analyzed, it was observed that there were also differences in the scores in other components of the rubric in terms of lower and/or higher rankings (see Table 11).

Table 11

Comparison between the Pre and Post-test for all the Scores Assigned by Rater # 4

Component	**Negative Ranks***	**Positive Ranks****	**Ties*****
Fluency and Pronunciation	**5**	**0**	**5**
Vocabulary	**7**	**1**	**2**
Grammatical Range and Accuracy	**2**	**2**	**6**
Task Completion	**3**	**2**	**5**
Comprehension	**2**	**2**	**6**
Total Score	**6**	**2**	**2**
TOTAL	**25**	**9**	**26**

* post-test scores < pre-test scores
** post-test scores > pre-test scores
*** post-test scores = pre-test scores

As seen in Table 11, while Rater # 4 assigned the same scores in nearly half of the scorings, a great majority of her scorings ranked lower in the post-test whereas she assigned only nine higher scores in the post-test. Given the significant differences in *Fluency and Pronunciation* and *Vocabulary* and the high frequency of negative

[22] Rater # 4 reported that she was familiar with two students, Student # 6 and Student # 10, so the scores assigned for these students were not included in data analysis.

ranks, the ranks in *Total Scores* should be analyzed in relation to the proficiency levels of the students (see Table 12).

Table 12

The Frequency of the Ranks in the Total Scores Assigned by Rater # 4[23] *for Each Level*

Levels[24]	**Negative Ranks ***	**Positive Ranks****	**Ties*****
D Levels	**2**	**2**	**0**
C Levels	**2**	**0**	**1**
B Levels	**2**	**0**	**1**
TOTAL	**6**	**2**	**2**

* post-test scores < pre-test scores
** post-test scores > pre-test scores
*** post-test scores = pre-test scores

As Table 12 shows, Rater # 4 was more severe towards C and B levels in the post-test *Total Scores* while she assigned both higher and lower scores equally for D level students. In order to examine whether Rater # 4 referred to the proficiency levels of students while assigning scores, especially in the components where significant differences were observed, the verbal reports of Rater # 4 were analyzed. Figure 9 shows some extracts from the pre and post-test verbal reports of Rater # 4 in relation to the scores she assigned.

[23] Rater # 4 reported that she was familiar with two students, Student # 6 and Student # 10, so the scores assigned for these students were not included in data analysis.

[24] D/C/B Levels (D: the lowest, B: the highest)

Student No	Student Level [25]	Partner's No & Level[26]	Component of the rubric	Pre-test score & comment	Post-test score & comment	Pre-test Total Score	Post-test Total Score
11	B	12- D	*Vocabulary*	(3) *Limited range in the first task. No problem that affected communication.*	(2) *Not inappropriate, but limited.*	(14)	(13)
			Total Score	(14) *He was better in the second task. They are also influenced from each other, by the structures and the vocabulary they used.*	(13) *B level student was not as good as I expected. I did not see a big difference between them. I think he should also have studied the 1st book** (* 1st book: the one for lower levels).		
12	D	11- B	*Fluency and Pronunciation*	(3) *Good fluency, no big mistake for pronunciation*	(2) *Had lots of mistakes, not fluent*	(15)	(11)
			Total Score	(15) *She was better in the first task. They are also influenced from each other, by the structures and the vocabulary they used.*	(11) *I did not see a big difference between them. D level student needs practice a lot.*	(15)	(11)

Figure 9. Examples of assigned scores and verbal reports by Rater # 9.

[25] D/C/B Levels (D: the lowest, B: the highest)

[26] This information is provided since some raters referred to the level of the students' partners and assigned scores by comparing them.

As seen in Figure 9, in the verbal reports of Rater # 4 provided for the performance of 10 students, there were two explicit references to the levels of the students, for Student # 11 who is a B level student, and for Student # 2, a D level student. Apart from these, there were no explicit references to the students' levels. However, it was observed that while assigning scores in the post-test for Student # 1, a D level student and Student # 2, a C level student, she stated *"First of all, so as not to repeat myself, I should mention that, in general, pronunciation is the basic problem for our students. Vocabulary range and use is also a big problem maybe because they think Turkish and they try to translate what they think."* Given these comments, as the results of Wilcoxon Signed Ranks Test indicated a significant difference in these components, the rater mostly assigned lower scores to the post-test *Fluency and Pronunciation* and *Vocabulary,* and as a result, the post-test *Total Scores* ranked lower. However, with the limited number of the references to the levels of the students, it is not possible to draw any conclusions about how different proficiency levels received attention from Rater # 4.

Data analysis for Rater # 6.

The results indicated a statistically significant difference between the pre and post-test scores assigned by Rater # 6[27] to the *Task Completion* (z = -2.333, p = .020). A further analysis on all the scores assigned by Rater # 6 revealed that there were also differences in the scores in the other components of the rubric in terms of lower and/or higher rankings (see Table 13).

[27] Rater # 6 reported that she was familiar with Student # 7, so the scores assigned for this student were not included in data analysis.

Table 13

Comparison between the Pre and Post-test for all the Scores Assigned by Rater # 6[28]

Component	**Negative Ranks***	**Positive Ranks****	**Ties*****
Fluency and Pronunciation	**2**	**3**	**6**
Vocabulary	**1**	**2**	**8**
Grammatical Range and Accuracy	**4**	**2**	**5**
Task Completion	**0**	**6**	**5**
Comprehension	**0**	**3**	**8**
Total Score	**2**	**4**	**5**
TOTAL	**9**	**20**	**37**

* post-test scores < pre-test scores
** post-test scores > pre-test scores
*** post-test scores = pre-test scores

As Table 13 shows, Rater # 6 was consistent in most of her scorings whereas she also assigned higher rankings in the 20 of her scorings. Only nine of her scorings ranked lower. Given the significant difference in *Task Completion* scores and high frequency of positive ranks, how different proficiency levels received different attention from Rater # 6 was analyzed (see Table 14).

[28] Rater # 6 reported that she was familiar with Student # 7, so the scores assigned for this student were not included in data analysis.

Table 14

The Frequency of the Ranks in the Total Scores Assigned by Rater # 6 for Each Level

Levels[29]	**Negative Ranks ***	**Positive Ranks****	**Ties*****
D Levels	**1**	**2**	**2**
C Levels	**0**	**2**	**0**
B Levels	**1**	**0**	**3**
TOTAL	**2**	**4**	**5**

* post-test scores < pre-test scores
** post-test scores > pre-test scores
*** post-test scores = pre-test scores

As seen in Table 14, Rater # 6 was consistent while assigning final scores to five students. However, when compared to B level students, she was more lenient while assessing the performances of the D and C level – the lower - levels students. In order to understand whether this change is a random or standard error of measurement, the verbal reports of Rater # 6 should be analyzed in line with the significant difference in *Task Completion* and high frequency of positive ranks. During think aloud protocols, the rater referred to the proficiency levels of the students explicitly while assigning scores for Student # 9, Student # 10, Student # 11, and Student # 12. Figure 10 shows some extracts from the pre and post-test verbal reports of Rater # 6 in relation to the scores she assigned.

[29] D/C/B Levels (D: the lowest, B: the highest)

Student No	Student Level[30]	Partner's No & Level[31]	Component of the rubric	Pre-test score & comment	Post-test score & comment	Pre-test Total Score	Post-test Total Score
9	D	10- B	*Total Score*	(19) *He was very successful and fluent, he used good sentence structures.*	(16) *For a D level student, he was successful and fluent although he had some pronunciation mistakes.*	(19)	(16)
10	B	9 - D	*Grammatical Range and Accuracy*	(2) *The student had few sentences and some mistakes.*	(3) *It could be better. She is not like a B level student.*	(11)	(11)
11	B	12 - D	*Fluency and Pronunciation*	(4) *He had some pauses, I liked that he used some expressions such as actually, especially, I mean, and his pronunciation was good.*	(3) *He answered the teacher's questions, but could give long answers for the topic music, it was an easy topic, he gave short answers. In the second task, he asked good questions to his friend. He spoke/performed well, but as B level student, he could do better.*	(18)	(17)
12	D	11 - B	*Vocabulary*	(3) *Cough, headache. Good appropriate vocabulary.*	(4) *Good vocabulary such as get stressed, cough, it was good considering she is a D Level student.*	(15)	(19)
			Task Completion	(3) *She did not understand her partner's questions and could not ask good relevant questions.*	(4) *She answered her partner's questions, but did not understand only one question, she could have asked more questions.*		

Figure 10. Examples of assigned scores and verbal reports by Rater # 6.

[30] D/C/B Levels (D: the lowest, B: the highest)

[31] This information is provided since some raters referred to the level of the students' partners and assigned scores by comparing them.

As a result, although there was a statistically significant difference in the scores assigned to *Task Completion*, there were also differences between the pre and post-test scores, especially in the *Fluency and Pronunciation* and *Grammatical Range and Accuracy* scores assigned for individual students. As seen in Figure 10, the references to the levels in the verbal reports of Rater # 6 revealed that the knowledge of the students' proficiency levels influenced her scorings. However, the rater was not consistent in assigning higher and/or lower scores to a specific proficiency level students.

Data analysis for Rater # 8.

The results indicated a statistically significant difference between the pre and post-test scores assigned by Rater # 8[32] to the *Grammatical Range and Accuracy* (z = -2.236, p = .025) and *Comprehension* (z = -2.000, p = .046). However, further analysis on the assigned scores revealed that there were also higher and/or lower rankings in the other components of the rubric (see Table 15).

Table 15

Comparison between the Pre and Post-test for all the Scores Assigned by Rater # 8

Component	**Negative Ranks***	**Positive Ranks****	**Ties*****
Fluency and Pronunciation	**3**	**2**	**6**
Vocabulary	**1**	**2**	**8**
Grammatical Range and Accuracy	**5**	**0**	**6**
Task Completion	**1**	**3**	**7**
Comprehension	**4**	**0**	**7**
Total Score	**6**	**1**	**4**
TOTAL	**20**	**8**	**38**

* post-test scores < pre-test scores
** post-test scores > pre-test scores
*** post-test scores = pre-test scores

[32] Rater # 8 reported that he was familiar with Student # 7, so the scores assigned for this student were not included in data analysis.

As seen in Table 15, Rater # 8 was mostly consistent in the scores she assigned; however, although there were statistically significant differences in the *Grammatical Range and Accuracy* and *Comprehension*, negatively ranked scores were also high in the *Total Scores*. In order to examine how Rater # 8 assigned scores for different proficiency level students, the frequency of ranks in the *Total Scores* assigned by the rater for each level should be analyzed (see Table 16).

Table 16

The Frequency of the Ranks in the Total Scores Assigned by Rater # 8[33] *for Each Level*

Levels[34]	**Negative Ranks ***	**Positive Ranks****	**Ties*****
D Levels	**3**	**0**	**1**
C Levels	**2**	**0**	**1**
B Levels	**1**	**1**	**2**
TOTAL	**6**	**1**	**4**

* post-test scores < pre-test scores
** post-test scores > pre-test scores
*** post-test scores = pre-test scores

As seen in Table 16, while the *Total Scores* of four students did not change in the post-test, six students' scores ranked lower, and one student's score ranked higher than their pre-test *Total Scores*. While B level students' scores ranked differently, Rater # 8 was more severe while assigning *Total Scores* for the D and C level students. The rater's verbal reports were analyzed to understand whether the knowledge of the students' proficiency levels influenced the rater's judgment. Figure 11 shows some extracts from the pre and post-test verbal reports of Rater # 8 in relation to the scores he assigned.

[33] Rater # 8 reported that he was familiar with Student # 7, so the scores assigned for this student were not included in data analysis.

[34] D/C/B Levels (D: the lowest, B: the highest)

Student No	Student Level[35]	Partner's No & Level[36]	Component of the rubric	Pre-test score & comment	Post-test score & comment	Pre-test Total Score	Post-test Total Score
9	D	10- B	*Vocabulary*	(4) *Appropriate terms, talked about caves.*	(4) *Appropriate terms and range such as caves. For a D level, very good.*	(18)	(16)
			Grammatical Range and Accuracy	(4) *No mistakes that obscured meaning*	(3) *Some mistakes that did not obscure meaning.*		
11	B	12 - D	*Total Score*	(16) *His partner was a little better than him especially in the first task, so she got 2 points higher than him.*	(18) *He was more fluent, enthusiastic. We should also consider that this student is a B level student, and the other one is a D level student.*	(16)	(18)
			Task Completion	(3) *Limited details, especially the second task was not like a dialog, but they were not irrelevant.*	(4) *Good details in both tasks, performed well, especially in the second task.*		
12	D	11 - B	*Total Score*	(18) *She was a little more successful than her partner especially in the first task, so she got 2 points higher than her partner*	(16) *Her partner was more fluent, enthusiastic. Both students were successful. We should also consider that this student is a D level student, and the other one is a B level student.*	(18)	(16)
			Vocabulary	(4) *She used appropriate words in the first task such as smoking.*	(3) *Not very detailed, but she used appropriate terms.*		

Figure 11. Examples of the assigned scores and verbal reports by Rater # 8.

[35] D/C/B Levels (D: the lowest, B: the highest)

[36] This information is provided since some raters referred to the level of the students' partners and assigned scores by comparing them.

As a result, although there were statistically significant differences in the scores assigned to the *Grammatical Range and Accuracy* and *Comprehension*, there were also differences between the pre and post-test scores, especially in the *Fluency and Pronunciation* and *Total Scores* assigned for individual students. As seen in Figure 11, Rater # 8 referred to the levels of three students, two of them being D level students. As Table 16 shows, while the D and C level students received mostly lower ranks in their *Total Scores,* in general, the rater was not consistent in assigning higher or lower scores to the B level students.

Data analysis for Rater # 14.

The findings indicated a statistically significant difference between the pre and post-test *Vocabulary* (z = -2.236, p = .025) scores assigned by Rater # 14[37]. However, when all the scores assigned by Rater # 14 were examined, some inconsistencies were also observed in the scores assigned for the other components of the rubric in terms of lower and/or higher rankings (see Table 17).

Table 17

Comparison between the Pre and Post-test for all the Scores Assigned by Rater # 14

Component	**Negative Ranks***	**Positive Ranks****	**Ties*****
Fluency and Pronunciation	**2**	**0**	**9**
Vocabulary	**5**	**0**	**6**
Grammatical Range and Accuracy	**3**	**0**	**8**
Task Completion	**1**	**2**	**8**
Comprehension	**1**	**2**	**8**
Total Score	**6**	**1**	**4**
TOTAL	**18**	**5**	**43**

* post-test scores < pre-test scores
** post-test scores > pre-test scores
*** post-test scores = pre-test scores

[37] Rater # 14 reported that she was familiar with Student # 10, so the scores assigned for this student were not included in data analysis.

As seen in Table 17, Rater # 14 was mostly consistent in her scorings by assigning the same scores. However, out of 11 students, six students' post-test *Total Scores* ranked lower than their pre-test scores while one student received favorable rankings. In order to analyze whether Rater # 14 assigned lower rankings in the *Total Scores* of a group of students with the same proficiency levels, the frequency of the ranks in the *Total Scores* assigned for each levels was analyzed (see Table 18).

Table 18

The Frequency of the Ranks in the Total Scores Assigned by Rater # 14[38] *for Each Level*

Levels[39]	**Negative Ranks ***	**Positive Ranks****	**Ties*****
D Levels	**1**	**1**	**3**
C Levels	**2**	**0**	**1**
B Levels	**3**	**0**	**0**
TOTAL	**6**	**1**	**4**

* post-test scores < pre-test scores
** post-test scores > pre-test scores
*** post-test scores = pre-test scores

Table 18 shows that Rater # 14 was more severe in her post-test *Total Score* scorings, especially for B and C level students. However, she was more consistent in her scorings for D level students although lower and higher rankings were observed. Given the significant difference between the pre and post-test *Vocabulary* scorings and lower ranks in the *Total Scores* of six out of 12 students, verbal reports by Rater # 14 were analyzed to find out how she perceived the performances of the students with and without the information of their proficiency levels while assigning scores and whether she referred to the proficiency levels of students. It was observed that the rater referred to the proficiency level of only one student. Figure 12 shows some extracts from her verbal reports in relation to the scores she assigned.

[38] Rater # 14 reported that she was familiar with Student # 10, so the scores assigned for this student were not included in data analysis.
[39] D/C/B Levels (D: the lowest, B: the highest)

Student No	Student Level[40]	Partner's No & Level[41]	Component of the rubric	Pre-test score & comment	Post-test score & comment	Pre-test Total Score	Post-test Total Score
12	D	11 – B	*Total Score*	(17) *They were both good, they were not very fluent, they did not speak comprehensively, but they are in the production phase.*	(18) *She was successful considering she is a D level student. She had good sentences and used appropriate vocabulary. They were not bad, they were fair average students.*	(17)	(18)
			Comprehension	(3) *I assigned 3 for the same reasons with her partner. They are good, but could be better.*	(4) *No problem, understood what is said*		

Figure 12. Examples of the assigned scores and verbal reports by Rater # 14.

As Figure 12 presents, although there was only one explicit reference to the level of a student, it was observed that the lower scores in *Vocabulary* were mostly assigned to higher proficiency levels. Out of five students who had a lower ranking in the post-test *Vocabulary*, there was only one D level student who was assigned a lower *Vocabulary* score in the post-test. Apart from the more severe rankings in the *Vocabulary* scores of higher proficiency level students, no other conclusions can be drawn from the available data.

Raters who did not refer to the proficiency levels of the students.

Although significant differences were observed in their scorings, and there were relatively higher and/or lower rankings in the post-test

[40] D/C/B Levels (D: the lowest, B: the highest)

[41] This information is provided since some raters referred to the level of the students' partners and assigned scores by comparing them.

Total Scores than the pre-test *Total Scores*, when their verbal reports were analyzed, it was found that two raters, Rater # 11 and Rater # 15, did not explicitly or implicitly refer to the proficiency levels of the students while assigning scores. However, it should be considered that similar to other raters, the information about the students' levels were given to these raters both in oral and written format.

Data analysis for Rater # 11.

The findings indicated a statistically significant difference between the pre and post-test scores assigned by Rater # 11[42] to the *Grammatical Range and Accuracy* (z = -2.236, p = .025). Further analysis of the scores assigned to the individual students indicated that the rater also assigned lower and/or higher scores to some students in other components of the rubric (see Table 19).

Table 19

Comparison between the Pre and Post-test for all the Scores Assigned by Rater # 11

Component	**Negative Ranks***	**Positive Ranks****	**Ties*****
Fluency and Pronunciation	**5**	**2**	**5**
Vocabulary	**3**	**2**	**7**
Grammatical Range and Accuracy	**5**	**0**	**7**
Task Completion	**4**	**1**	**7**
Comprehension	**2**	**3**	**7**
Total Score	**8**	**2**	**2**
TOTAL	**27**	**10**	**35**

* post-test scores < pre-test scores
** post-test scores > pre-test scores
*** post-test scores = pre-test scores

As seen in Table 19, Rater # 11 assigned the same scores in a great majority of the scorings while 27 of her scorings ranked lower and 10 of her scorings ranked higher.

[42] Rater # 11 reported that he was not familiar with any of the 12 students.

In the scores other than the *Grammatical Range and Accuracy* in which a significant difference was observed, a great majority of the *Total Scores* changed in the post-test with a lower ranking of the scores of eight students. In order to analyze how different proficiency levels received different attention from Rater # 11, the frequency of ranks in the *Total Scores* assigned for each level were examined (see Table 20).

Table 20

The Frequency of the Ranks in the Total Scores Assigned by Rater # 11 for Each Level

Levels[43]	**Negative Ranks ***	**Positive Ranks****	**Ties*****
D Levels	**2**	**2**	**1**
C Levels	**3**	**0**	**0**
B Levels	**3**	**0**	**1**
TOTAL	**8**	**2**	**2**

* post-test scores < pre-test scores
** post-test scores > pre-test scores
*** post-test scores = pre-test scores

As Table 20 shows, eight students out of 12 received negative ranks in the post-test *Total Sores.* While Rater # 11 assigned both negative and positive ranks for D level students, he was more severe in his scorings for C and B level students. When Rater # 11's verbal reports were analyzed, it was observed that he did not refer to the students' proficiency levels while assigning scores at any point explicitly, or even implicitly. To illustrate his verbal reports, Figure 13 shows some extracts from the pre and post-test verbal reports of Rater # 11 in relation to the scores he assigned.

[43] D/C/B Levels (D: the lowest, B: the highest)

Student No	Student Level[44]	Partner's No & Level[45]	Component of the rubric	Pre-test score & comment	Post-test score & comment	Pre-test Total Score	Post-test Total Score
11	B	12 - D	*Grammatical Range and Accuracy*	(4) *Good. No big mistakes, some minor errors.*	(3) *Some minor errors, but they did not obscure meaning*	(18)	(16)
12	D	11 - B	*Grammatical Range and Accuracy*	(3) *The student had some minor mistakes that did not obscure meaning.*	(4) *Good use of when and if clauses which our students usually have problems, but the student had some problems in sentence structures.*	(13)	(13)

Figure 13. Examples of the assigned scores and verbal reports by Rater # 11.

As a result, it is impossible to draw any conclusions about why Rater # 11 assigned statistically different scores to five students' *Grammatical Range and Accuracy* and reported different opinions about the students' performances. More importantly, no conclusion from this data can be drawn about why eight students received lower ranks and two students were assigned more favorable scores while the scores of two students did not change.

Data analysis for Rater # 15.

The findings indicated a statistically significant difference between the pre and post-test *Task Completion* scores (z = -2.646, p = .008)

[44] D/C/B Levels (D: the lowest, B: the highest)

[45] This information is provided since some raters referred to the level of the students' partners and assigned scores by comparing them.

assigned by Rater # 15[46]. Further analysis of all the scores assigned by the rater indicated that he also assigned lower and/or higher scores to some students in other components of the rubric (see Table 21).

Table 21

Comparison between the Pre and Post-test for all the Scores Assigned by Rater # 15

Component	**Negative Ranks***	**Positive Ranks****	**Ties*****
Fluency and Pronunciation	**2**	**4**	**6**
Vocabulary	**3**	**4**	**5**
Grammatical Range and Accuracy	**1**	**3**	**8**
Task Completion	**0**	**7**	**5**
Comprehension	**3**	**4**	**5**
Total Score	**2**	**8**	**2**
TOTAL	**11**	**30**	**31**

* post-test scores < pre-test scores
** post-test scores > pre-test scores
*** post-test scores = pre-test scores

As seen in Table 21, although Rater # 15 assigned equal scores in many of the pre and post-test scorings, positive ranks were also very high in all the components of the rubric, especially in the *Task Completion* and *Total Scores*. In order to examine how different scorings each level of students received in their *Total Scores*, the frequency of the ranks in the *Total Scores* assigned for each level was analyzed (see Table 22).

[46] Rater # 15 reported that he was not familiar with any of the students.

Table 22

The Frequency of the Ranks in the Total Scores Assigned by Rater # 15 for Each Level

Levels[47]	**Negative Ranks ***	**Positive Ranks****	**Ties*****
D Levels	**1**	**2**	**2**
C Levels	**0**	**3**	**0**
B Levels	**1**	**3**	**0**
TOTAL	**2**	**8**	**2**

* post-test scores < pre-test scores
** post-test scores > pre-test scores
*** post-test scores = pre-test scores

Table 22 shows that Rater # 15 was more lenient while assigning the post-test *Total Scores* for eight students from three different proficiency levels. Especially all the C level students and B level students except for one received more favorable rankings in their *Total Scores*. Given the highly significant difference in the *Task Completion* scores and the high frequency of positive ranks especially in the *Total Scores,* the verbal reports of Rater # 15 were analyzed to observe whether the higher scores in the post-test were assigned due to the influence of the rater's knowledge of the students' proficiency levels. However, in his post-test verbal reports, the rater did not ever mention the proficiency levels of the students. Further analysis on his verbal reports revealed that the inconsistencies were basically in the *Task Completion* in which a significant difference was observed. Since while assigning scores to *Task Completion,* the two tasks are considered, the inconsistencies were mostly about how Rater # 15 assessed the students' misunderstanding the topics, task difficulty, limited sentence production, and the effects of the other candidate's poor performance in the second task. Figure 14 shows some extracts from the pre and post-test verbal reports of Rater # 15 in relation to the scores he assigned.

[47] D/C/B Levels (D: the lowest, B: the highest)

Student No	Student Level [48]	Partner's No & Level[49]	Component of the rubric	Pre-test score & comment	Post-test score & comment	Pre-test Total Score	Post-test Total Score
2	C	1 - D	*Task Completion*	(2) *The first task was difficult. She started appropriately, but could not continue. In the second task, she usually continued the dialog, but while asking questions, he did not ask relevant questions.*	(3) *I don't think there was a problem. Especially the topic of the first task was difficult. Although she did not deal with the topic comprehensively, she did her best.* *In the second task, she tried to interact, communicate, but her partner was not active, enthusiastic, so she had some problems here.*	(13)	(15)
10	B	9 - D	*Task Completion*	(1) *She did not speak almost at all in the first task and needed frequent encouragement to speak, so she did not complete the task successfully. Similarly in the second task, her partner spoke mostly. When it was her turn to speak, she did not produce many sentences.*	(2) *She did not produce many sentences in both tasks, so it was very difficult to assess her performance. She produced short answers, I don't think that she completed the tasks successfully. In the first task, she usually needed the teacher's guidance, but she gave short answers* *In the second task, while the other student was talking, she did not try to interrupt, take turn. Although her partner talked too long, she listened until the end, and she produced only short sentences.*	(9)	(12)

Figure 14. Examples of the assigned scores and verbal reports by Rater # 15.

[48] D/C/B Levels (D: the lowest, B: the highest)

[49] This information is provided since some raters referred to the level of the students' partners and assigned scores by comparing them.

As seen in Figure 14, Rater # 15 did not refer to the proficiency levels of the students while he was assigning scores in the post-test. However, task difficulty and the performance of the candidate's partner were the themes that emerged frequently in his verbal reports. As a result, considering the available data, it was concluded that it is impossible to draw any further conclusions apart from these emerged themes.

When the frequency of the ranks in the *Total Scores* assigned by these eight raters were analyzed, it was found that for each level, the raters mostly assigned lower or higher scores in the post-test (see Table 23).

Table 23

The Frequency of the Ranks in the Total Scores Assigned by Eight Raters[50] for Each Level

Levels[51]	**Negative Ranks ***	**Positive Ranks****	**Ties*****
D Levels	**12**	**13**	**11**
C Levels	**9**	**9**	**4**
B Levels	**14**	**9**	**7**
TOTAL	**35**	**31**	**22**

* post-test scores < pre-test scores
** post-test scores > pre-test scores
*** post-test scores = pre-test scores

As seen in Table 23, in general, while the number of negative and positive ranks were almost equal in the *Total Scores* assigned for D and C level students, the number of negative ranks assigned for B level students were higher than the positive ranks. D level - the lowest proficiency level- students received a slightly more favorable rankings in their scorings while B level – the highest proficiency level- students received more severe scorings in their post-test *Total Scores*.

[50] The raters who had significant differences between their pre and post-test scorings.
[51] D/C/B Levels (D: the lowest, B: the highest)

Data analysis for the raters with no statistically significant difference between their scorings.

As presented in Table 2, seven raters, Rater # 2, Rater # 5, Rater # 7, Rater # 9, Rater # 10, Rater # 12, and Rater # 13, showed statistically no significant different scoring behavior in the post-test. However, when the *Total Scores* assigned by these seven raters were analyzed in terms of positive and negative ranks, the results indicated that similar to the eight raters who had significant differences in their scorings, these raters also assigned higher and/or lower scores for some students in the post-test, and there are similarities between the percentages of the rankings assigned by these seven raters and the eight raters with significant differences (see Table 4 and Table 5). In this section, the data gathered from these seven raters' scorings and verbal reports will be presented in two parts. First, the data from the five raters who referred to the proficiency levels of the students in their verbal reports will be analyzed. Then, the data from the two raters who did not mention the levels of the students will be introduced.

Raters who referred to the proficiency levels of the students.

As shown in Table 2, there was no significant difference between the pre and post-test scores assigned by Rater # 2[52], Rater # 7[53], Rater # 9[54], Rater # 10[55], and Rater # 12[56]. However, further analysis of all the scores assigned to each student by these raters indicated that they assigned lower and/or higher scores to some students in all components of the rubric (see Table 24).

[52] Rater # 2 reported that she was not familiar with any of the students.

[53] Rater # 7 reported that he was familiar with Student # 4, so the scores assigned for this student was not included in data analysis.

[54] Rater # 9 reported that she was not familiar with any of the students.

[55] Rater # 10 reported that he was not familiar with any of the students.

[56] Rater # 12 reported that she was not familiar with any of the students.

Table 24

Comparison between the Pre and Post-test for all the Scores Assigned by the Raters[57]

Rater	***Rater # 2***			***Rater # 7***			***Rater # 9***			***Rater # 10***			***Rater # 12***			***Total Ranks by Five Raters***		
Component	**Negative Ranks***	**Positive Ranks****	**Ties*****	**Negative Ranks***	**Positive Ranks****	**Ties*****	**Negative Ranks***	**Positive Ranks****	**Ties*****	**Negative Ranks***	**Positive Ranks****	**Ties*****	**Negative Ranks***	**Positive Ranks****	**Ties*****	**Negative Ranks***	**Positive Ranks****	**Ties*****
Fluency and Pronunciation	**3**	**1**	**8**	**2**	**3**	**6**	**4**	**1**	**7**	**2**	**1**	**9**	**4**	**0**	**8**	**15**	**6**	**38**
Vocabulary	**4**	**2**	**6**	**5**	**2**	**4**	**2**	**3**	**7**	**2**	**0**	**10**	**1**	**1**	**10**	**14**	**8**	**37**
Grammatical Range and Accuracy	**5**	**2**	**5**	**1**	**3**	**7**	**1**	**0**	**11**	**1**	**2**	**9**	**3**	**3**	**6**	**11**	**10**	**38**
Task Completion	**3**	**2**	**7**	**2**	**3**	**6**	**5**	**2**	**5**	**1**	**5**	**6**	**3**	**1**	**8**	**14**	**13**	**32**
Comprehension	**3**	**5**	**4**	**5**	**3**	**3**	**3**	**2**	**7**	**3**	**2**	**7**	**1**	**3**	**8**	**15**	**15**	**29**
Total Score	**6**	**4**	**2**	**4**	**3**	**4**	**6**	**3**	**3**	**4**	**5**	**3**	**6**	**4**	**2**	**26**	**19**	**14**
TOTAL	**24**	**16**	**32**	**19**	**17**	**30**	**21**	**11**	**40**	**13**	**15**	**44**	**18**	**12**	**42**	**95**	**71**	**188**

* post-test scores < pre-test scores
** post-test scores > pre-test scores
*** post-test scores = pre-test scores

[57] The raters without a significant difference but with reference to the levels

As seen in Table 24, the high frequency of ties indicated that these five raters were mostly consistent within themselves in their scorings. However, when the scores assigned to each component of the rubric were analyzed, it was observed that each of these raters behaved differently, especially when the number of higher and lower ranks assigned to the *Total Scores* were considered. When all the scores assigned by Rater # 2 were analyzed, Rater # 2 was more lenient while assigning scores for *Comprehension,* but she assigned lower rankings in *Vocabulary, Grammatical Range and Accuracy,* and *Total Score* more frequently than she did in the other components of the rubric. Out of 12 students, six students' post-test *Total Scores* assigned by Rater # 2 ranked lower than their pre-test scores while four students received favorable rankings and two students were assigned the same scores.

As for Rater # 7, despite the high frequency of equal scores and the existence of some positive scores, Rater # 7 assigned lower rankings in *Vocabulary, Comprehension,* and *Total Score* more frequently than he did in the other components of the rubric. Out of 11 students' post-test *Total Scores,* it was examined that four students received lower ranks while three students were assigned higher scores, and the scores of four students did not change.

Although Rater # 9 was mostly consistent in her scorings and there was almost absolute agreement on the scores she assigned to *Grammatical Range and Accuracy,* the results indicated that she also assigned lower and/or higher ranks to some students in different components of the rubric. The highest difference between the positive and negative ranks was observed in the *Fluency and Pronunciation, Task Completion,* and *Total Scores*. Out of 12 students, the *Total Scores* of three students did not change while three students received higher scores and six students were assigned lower ranks.

Similarly, Rater # 10 was also mostly consistent while assigning scores. Despite the high number of equal scores and the existence of negative ranks assigned to some students in all components of the rubric, positive ranks were more frequently observed than the negative ranks in *Task Completion* and *Total Scores*. Out of 12 students, the *Total Scores* of three students did not change while four students received lower scores and five students were assessed more favorably.

Last but not least, Rater # 12 also presented some inconsistencies in her scorings in terms of lower and/or higher scores assigned in different components of the rubric. She assigned more positive scores than lower scores in *Comprehension,* but the negative ranks were more than the positive ranks in *Fluency and Pronunciation, Task Completion, and Total Scores.* Moreover, the equal scores were more frequent than the negative and positive ranks in each component except for the *Total Scores.* Out of 12 students, while the *Total Scores* of two students did not change, four students received more favorable scores, but six students were assigned lower *Total Scores* in the post-test. As a result, although the results indicated no significant difference between the scores assigned by these raters, as shown in Table 24, a majority of the students who were assessed received lower *Total Scores*, but the number of positive ranks were also higher than the equal scores. Given the high frequency of lower and higher scores in the post-test *Total Scores* assigned to several students, further analysis was conducted in order to see how different rankings each group of students, who were grouped according to their proficiency levels, received in their *Total Scores* assigned by these five raters (see Table 25).

Table 25

The Frequency of the Ranks in the Total Scores Assigned for Each Level by the Raters[58]

Rater	***Rater # 2***			***Rater # 7***			***Rater # 9***			***Rater # 10***			***Rater # 12***			***Total Ranks by 5 Raters***		
Levels[59]	**Negative Ranks***	**Positive Ranks****	**Ties*****	**Negative Ranks***	**Positive Ranks****	**Ties*****	**Negative Ranks***	**Positive Ranks****	**Ties*****	**Negative Ranks***	**Positive Ranks****	**Ties*****	**Negative Ranks***	**Positive Ranks****	**Ties*****	**Negative Ranks***	**Positive Ranks****	**Ties*****
D Levels	**4**	**1**	**0**	**1**	**1**	**3**	**3**	**1**	**1**	**2**	**2**	**1**	**3**	**1**	**1**	**13**	**6**	**6**
C Levels	**1**	**1**	**1**	**2**	**1**	**0**	**1**	**1**	**1**	**1**	**2**	**0**	**1**	**2**	**0**	**6**	**7**	**2**
B Levels	**1**	**2**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**2**	**2**	**1**	**1**	**7**	**6**	**6**
TOTAL	**6**	**4**	**2**	**4**	**3**	**4**	**6**	**3**	**3**	**4**	**5**	**3**	**6**	**4**	**2**	**26**	**19**	**14**

* post-test scores < pre-test scores
** post-test scores > pre-test scores
*** post-test scores = pre-test scores

As shown in Table 25, these five raters were not consistent among themselves while assigning lower and/or higher *Total Scores* in the post-test for each level of students. However, as mentioned before, a majority of the *Total Scores* ranked lower in the post-test. When the frequency of the ranks in the *Total Scores* assigned by each rater for each level was analyzed, it was observed that while some raters were more lenient towards the higher level students, most of the lower level students received more severe scorings. Rater # 2 was more severe while assigning scores for six students in the post-test and more lenient towards four students. While Rater # 2 assigned both negative and positive ranks for C and B level students, she was more severe in her scorings for D level students. Rater # 7 was slightly more severe towards C level students; however, there was no strong pattern in the scores he assigned in terms of severity or leniency towards a specific level of students. Rater #

[58] The raters without a significant difference but with reference to the levels
[59] D/C/B Levels (D: the lowest, B: the highest)

9 mostly assigned lower scores in the post-test. The number of students who received lower scores were higher among the B and D level students. Three students out of five D level students got lower rankings in their post-test *Total Scores*. While Rater # 10’s scores for two out of three C level students increased, he assigned equally lower and higher *Total Scores* for the other levels. However, D and B level students received more negative ranks rather than positive ranks from Rater # 12, while, out of three C level students, the *Total Scores* of two students ranked higher. As a result, while three raters were more severe in their scorings for D level students by assigning lower scores for at least three of five D level students, the other two raters assigned equally lower and higher scores for D level students. While two raters assigned higher scores for two of three C level students in the post-test, the scores of one rater ranked lower for two C level students, and two raters assigned lower, equal, or higher scores for each student. In the scorings of four B level students, while one rater was more lenient for two students, two raters were more severe towards two students. When all the scores were considered, the results indicated that the number of lower, equal and higher scores assigned to B level students were almost the same, but the scores of C level students changed the most in terms of negative or positive ranks. Out of 15 scorings assigned for C level students, only two did not change. Moreover, half of the scorings assigned to the D level students ranked lower in the post-test. Further qualitative analysis of the verbal reports by these five raters revealed that they referred to the proficiency levels of some students while assessing the oral performances of the students. Figure 15 shows some extracts from the pre and post-test verbal reports of Rater # 2, Rater # 7, Rater # 9, Rater # 10, and Rater # 12 in relation to the scores they assigned.

Rater No	Student No	Student Level[60]	Partner's No & Level[61]	Component of the rubric	Pre-test score & comment	Post-test score & comment	Pre-test Total Score	Post-test Total Score
2	2	C	1 - D	*Vocabulary*	(4) *The student was very excited in the first task. The second task was very good, asked all the questions and used all the necessary words. She used connectors such as unfortunately.*	(2) *Although she is a C level student, she was very excited and had limited vocabulary range, the word "unfortunately" is the only the word range we can see.*	(18)	(8)
				Grammatical Range and Accuracy	(4) *She used "Should" unexpectedly, used present simple tense. It was good.*	(1) *She had lots of mistakes e.g., I like she, she don't. Grammar mistakes even in simple sentences, they obscured the meaning,*		
2	11	B	12 - D	*Total Score*	(18) *Both of them were successful in different areas. Student # 11 used good conversational strategies and expressions in the second task, but in the first task, he got help from the teacher. Student # 12 was successful in the first task in vocabulary and grammar use, but in the second task she did not ask many questions. For these reasons, I cut 2 points.*	(19) *When I see he is a B level student, honestly I have higher expectations, I am not sure if this is the right thing to do. Still, 19 is a good score.*	(18)	(19)
2	12	D	11 - B	*Total Score*		(14) *She had major problems in grammar, and she did not understand the second task.*	(18)	(14)

[60] D/C/B Levels (D: the lowest, B: the highest)

[61] This information is provided since some raters referred to the level of the students' partners and assigned scores by comparing them.

7	2	C	1 - D	*Vocabulary*	(2) *She could tell her ideas only by using adjectives.*	(4) *The student had adequate range for this level of student.*	(10)	(15)
				Grammatical Range and Accuracy	(1) *She formulated wrong sentences. The word order and word choice was wrong.*	(2) *Almost all the sentences were full of errors, and they obscured the meaning.*		
9	10	B	9 - D	*Total Score*	(13) *Her partner was better. She was less successful compared to her partner, but in pair work, it was obvious that this was a pair work, they asked questions to each other.*	(10) *Her partner continued the conversation although he was a D level student. She was passive although she was a B level student, she performed less successfully than her partner.*	(13)	(10)
				Task Completion	(3) *She understood the second task but could not express herself well, but expressed her ideas well in the first task, but also used short answers like do you want to get married: yes. Limited details.*	(1) *She could not speak in the first task at all. In the second task, she couldn't complete it successfully, either. She did not try much. Her performance was very poor.*		
9	12	D	11 - B	*Vocabulary*	(3) *She could have had more range, could have done better.*	(2) *Although she is a D level student, she could have more vocabulary range considering her level.*	(17)	(11)
10	12	D	11 - B	*Vocabulary*	(3) *She did not use sophisticated words, but did not have errors.*	(3) *Although she used similar basic words, I think she could accomplish what was expected of her in terms of vocabulary.*	(14)	(13)

						She did not have word errors and she used words appropriate to her level. She had problems in grammar, her vocabulary use was not very good, but not very bad.		
12	9	D	10 - B	*Vocabulary*	(4) *Good range. He used topic related words.*	(4) *According to his level, his vocabulary was very good.*	(19)	(18)
				Grammatical range and Accuracy	(4) *He made an error in only one sentence in the dialog, he said "I planning", maybe it is because while speaking he had mistakes.*	(3) *I heard errors in four sentences, and they were simple structures. Considering his level and the fact that I can ignore these errors, I cut only 1 point.*		
12	11	B	12 – D	*Vocabulary*	(4) *No problem, very good. He used the connectors effectively.*	(4) *He used appropriate words according to his level.*	(19)	(17)
				Task Completion	(4) *He continued the dialog, he was active. He helped and guided his partner.*	(3) *He was passive in the first task, but the second task was very good.*		

Figure 15. Examples of the assigned scores and verbal reports by Rater # 2, Rater # 7, Rater # 9, Rater # 10, and Rater # 12.

Although there was no significant difference between the pre and post-test scores assigned by Rater # 2, Rater # 7, Rater # 9, Rater # 10, and Rater # 12, as seen in Figure 15, when their verbal reports were analyzed in relation to the scores assigned to individual students, it was observed that these raters referred to the proficiency levels of the students and assigned lower or higher scores in the post-test. As discussed before, there was no pattern about how different attention each level received from the raters, but most of the raters referred to the proficiency levels of the same two students, Student # 11 and Student # 12 who were a B and a D level matched-pair. In other words, the highest proficiency level and the lowest proficiency level matched-pair received utmost attention from the raters.

Raters who did not refer to the proficiency levels of the students.

Although there was no significant difference between the pre and post-test scores of Rater # 5[62] and Rater # 13[63], further analysis of all the scores assigned to the individual students by these raters indicated that they assigned lower and/or higher scores in almost all components of the rubric (see Table 26).

[62] Rater # 5 reported that she was familiar with two students, Student # 4 and Student # 7, so the scores assigned for these students were not included in data analysis.

[63] Rater # 13 mentioned that she was familiar with Student # 12, so the scores assigned for this student were not included in data analysis.

Table 26

Comparison between the Pre and Post-test for all the Scores Assigned by the Raters[64]

Rater	***Rater # 5***			***Rater # 13***			***Total Ranks by Two Raters***		
Component	**Negative Ranks***	**Positive Ranks****	**Ties*****	**Negative Ranks***	**Positive Ranks****	**Ties*****	**Negative Ranks***	**Positive Ranks****	**Ties*****
Fluency and Pronunciation	**0**	**0**	**10**	**3**	**3**	**5**	**3**	**3**	**15**
Vocabulary	**1**	**3**	**6**	**2**	**2**	**7**	**3**	**5**	**13**
Grammatical Range and Accuracy	**3**	**2**	**5**	**2**	**3**	**6**	**5**	**5**	**11**
Task Completion	**2**	**4**	**4**	**3**	**1**	**7**	**5**	**5**	**11**
Comprehension	**2**	**2**	**6**	**2**	**5**	**4**	**4**	**7**	**10**
Total Score	**2**	**3**	**5**	**5**	**5**	**1**	**7**	**8**	**6**
TOTAL	**10**	**14**	**36**	**17**	**19**	**30**	**27**	**33**	**66**

* post-test scores < pre-test scores
** post-test scores > pre-test scores
*** post-test scores = pre-test scores

As shown in Table 26, despite the high frequency of equal scores, a great majority of the scorings ranked lower or higher in the post-test. While Rater # 5 assigned the same *Fluency and Pronunciation* scores for all the students in the pre and post-test, some students received lower or higher scores in the other components of the rubric. The *Total Scores* she assigned in the post-test did not change for five students, but she was more lenient towards three students and more severe towards two students. Similar to Rater # 5, Rater # 13 also assigned the same scores for some students in all components of the rubric. However, unlike Rater # 5, she assigned more favorable scores in *Comprehension,* and out of the scorings of 11 students, only

[64] The raters without a significant difference and no reference to the levels

the *Total Score* of one student did not change while five students received lower ranks and five students received higher scores. Moreover, for some students, the difference between the pre and post-test *Total Score* was very high such as 6 points when it was considered that the highest *Total Score* that can be assigned is 20 points. Table 27 presents the distribution of the ranks in the *Total Scores* assigned by these two raters according to the proficiency levels of the students.

Table 27

The Frequency of the Ranks in the Total Scores Assigned for Each Level by the Raters[65]

Rater	***Rater # 5***			***Rater # 13***			***Total Ranks by 2 Raters***		
Levels[66]	**Negative Ranks***	**Positive Ranks****	**Ties*****	**Negative Ranks***	**Positive Ranks****	**Ties*****	**Negative Ranks***	**Positive Ranks****	**Ties*****
D Levels	**2**	**1**	**1**	**2**	**1**	**1**	**4**	**2**	**2**
C Levels	**0**	**1**	**2**	**2**	**1**	**0**	**2**	**2**	**2**
B Levels	**0**	**1**	**2**	**1**	**3**	**0**	**1**	**4**	**2**
TOTAL	**2**	**3**	**5**	**5**	**5**	**1**	**7**	**8**	**6**

* post-test scores < pre-test scores
** post-test scores > pre-test scores
*** post-test scores = pre-test scores

As seen in Table 27, despite the existence of equal scorings in the *Total Scores* in three levels, the results indicated that the only two negative ranks assigned by Rater # 5 were observed in the scorings of two D level students. In other words, Rater # 5 was more severe towards the two D level students when the scores she assigned to the other levels were considered. In the scorings of Rater # 13, while the *Total Scores* of D and C level students ranked lower mostly in the post-test, three out of four B level students received higher scores in the post-test.

When their verbal reports were analyzed, it was found that two raters, Rater #

[65] The raters without a significant difference and no reference to the levels

[66] D/C/B Levels (D: the lowest, B: the highest)

5 and Rater # 13, did not implicitly or explicitly refer to the proficiency levels of the students while assigning scores. However, it should be noted that similar to the other raters, the information about the students' levels were provided to these raters both in oral and written format. Figure 16 presents some extracts from the pre and post-test verbal reports of Rater # 5 and Rater # 13 in relation to the scores they assigned.

Rater	Student No	Student Level [67]	Partner's No & Level[68]	Component of the rubric	Pre-test score & comment	Post-test score & comment	Pre-test Total Score	Post-test Total Score
5	1	D	2 - C	*Total Score*	(12) *I think she could not speak almost at all because of her anxiety. She was not fluent, she used limited vocabulary and had grammar mistakes.*	(10) *I think because of her anxiety, she had difficulty, she had no fluency, she had limited vocabulary range and grammar mistakes. There was a disrupted communication with her partner*	(12)	(10)
5	11	B	12 - D	*Total Score*	(15) *He could not express himself in the first task, limited range in vocabulary and grammar. He was better in the second task, more fluent.*	(19) *He was a successful student in all areas. Because of his hesitations/pauses in the first task, I cut 1 point from fluency, but he was successful in other areas.*	(15)	(19)
13	4	B	3 - B	*Grammatical Range and Accuracy*	(2) *The student had noticeable mistakes in the use of verb be, so*	(4) *The student had some minor mistakes while using verb be, but they did not*	(16)	(19)

[67] D/C/B Levels (D: the lowest, B: the highest)

[68] This information is provided since some raters referred to the level of the students' partners and assigned scores by comparing them.

					there was a problem in the sentence construction competency. She could not use passive voice correctly, but I do not consider it as an error because we did not teach it in our curriculum. In the second task, she had some verb tense mistakes, especially because she forgot to use verb be frequently in her sentences, I assign 2.	*obscure meaning. She used lots of complex sentences, both students used them and they made no mistake while using them. She used the connectors correctly.*		
13	10	B	9 - D	*Grammatical Range and Accuracy*	(1) *The student produced very few sentences, and not all of these sentences were true, and she had errors in grammar.*	(3) *The student used simple present tense while talking about last year for the topic of the second task: best vacation. I cannot say much about grammar in the first task because of the limited data. The student produced two or three sentences there.*	(5)	(11)

Figure 16. Examples of the assigned scores and verbal reports by Rater # 5 and Rater # 13.

Although there was no significant difference between each rater's pre and post-test scores and although these raters did not refer to the proficiency levels of the students during think-aloud protocols, the results indicated that there were inconsistencies in their verbal reports and scorings. As seen in Figure 16, although Rater # 5 reported almost exactly the same points for Student # 1, she assigned a lower *Total Score* in the post-test. However, while assigning scores for Student # 11, she was more favorable in her comments and scores. Although Rater # 13 mentioned the same problems in the pre and post-test, she assigned favorable scores for these B level students in the post-test. As a result, when the quantitative and qualitative data were analyzed, it was observed that although the difference between the pre and post-test scores was not statistically significant and although these raters did not refer to the levels of the students, it was observed that the raters assigned higher or lower scores to some students some of which were very high such as six points difference considering the fact that the videos used for this study were the samples taken from an oral interview exam which had been conducted as a part of a final proficiency exam.

The analysis of the verbal reports provided by the 15 raters revealed that while some raters referred to the proficiency levels of students, some did not mention it at all. Considering the references to the levels, it was observed that some raters' scores changed when they referred to the levels. While some raters were more severe in their scorings for higher proficiency levels, some were more lenient. Some raters assigned lower scores for lower proficiency levels while others were more favorable in their scorings for lower proficiency level students. However, there were also cases that the pre and post-test scorings of some raters were consistent although they mentioned the level of the student whose performance they assessed.

Conclusion

In this chapter, the descriptive statistics and the findings from the quantitative data were presented for each rater in relation to the qualitative data gathered from the think-aloud protocols. First, the overall quantitative data regarding the statistics of 15 raters' pre and post-test scorings were introduced. It was observed that in the scores assigned by eight raters, there were statistically significant differences between the pre and post-test treatment. The scores and the verbal reports by these raters were analyzed in the next section, and it was found that six raters who had significant difference between their pre and post-test scorings referred to the proficiency levels of students while two raters did not. Moreover, more in depth analysis of these verbal reports revealed that while some raters changed their scores when they referred to the levels of students, some raters were consistent in their scorings and comments. In terms of the leniency or severity towards the students with the same proficiency levels, each of these eight raters behaved differently, but more severity was observed in the B level students' *Total Scores* assigned by six raters. In the last section, quantitative and qualitative data gathered from the seven raters who did not have significant differences between their pre and post-test scorings were introduced. Although the overall findings from the quantitative data showed that there was no significant difference between their scorings, there were some inconsistencies in the pre and post-test scores of some students. The follow-up qualitative analysis for these cases demonstrated that five of these seven raters also referred to the proficiency levels of the students in the post-test while two raters did not mention the levels of the students at all. When the *Total Scores* assigned by these seven raters were analyzed considering the distribution of lower and/or higher scores assigned to the students grouped according to their proficiency levels, five raters assigned lower scores for more D level students. Although each group of these 15 raters behaved differently in their scorings for D and B level students, that is while one group of raters was more severe towards B level students, the other was more harsh in their scorings for D level students, the results indicated that the most frequent inconsistency was observed between the pre and post-test scorings for C level students. Out of 43 scorings for C level students, only eight did not change, but the number of negative and positive ranks were almost the same. Given the findings in this chapter, the discussion of the results will be presented in the following chapter, especially with a focus on how the data answer the research question of this study. Moreover, in addition to the discussion of the limitations and implications of the study, suggestions will be made for further research.

CHAPTER V: CONCLUSION

Introduction

The purpose of this quasi-experimental study was to investigate the effect of raters' prior knowledge of students' proficiency levels on their scorings of oral interview performances. In this respect, this study addressed the following research question:

- To what extent does raters' prior knowledge of students' proficiency level influence their assessment behaviors during oral interviews?

In this study with 15 raters from a state university in Turkey, two sets of instruments were employed: (a) the rating materials included video recordings, rating scale (see Appendix 1), and grading sheets (see Appendix 2 and 3), and (b) the data collection instruments were the scores and verbal reports provided by the raters (see Figure 4).

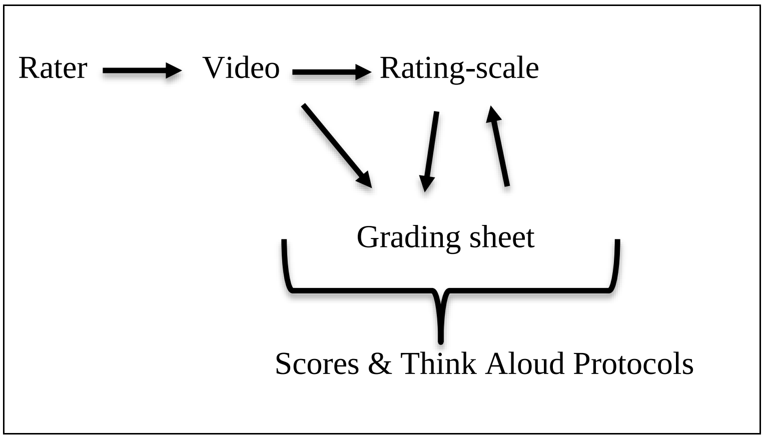

Figure 4. The interaction among the instruments during a scoring session conducted in this study.

This chapter consists of four main sections. In the first section, the findings emerged from this study will be discussed in relation to the similar studies conducted on rater effects. In the next section, the pedagogical implications will be introduced.

In the third section, the limitations of the study will be discussed, and in the final section, suggestions for further research will be presented.

Findings and Discussion

The Effects of the Raters' Prior Knowledge of Students' Proficiency Levels on their Assessment During Oral Interviews

The research question of the present study aimed to explore whether the raters' prior knowledge of students' proficiency levels is one of the rater effects that have an influence on their assessment behaviors during oral interviews. In this respect, first, quantitative data analysis was conducted to investigate whether there was a significant difference between the pre and post-test scores assigned by each rater. The rubric used in this study included five components which are *Fluency and Pronunciation, Vocabulary, Grammatical Range and Accuracy, Task Completion* and *Comprehension.* For each component, the lowest score that can be assigned is 1 point while the highest score is 4 points. As a *Total Score*, the raters can assign 5 points as the lowest score to a very poor performing student while the students with a successful performance can get up to 20 points. As a result, there were six scores assigned to one student by each rater both in the pre and post-test. The result of the data analysis revealed that there was a significant difference between the eight raters' pre and post-test scores assigned to different components of the rubric. Although only one rater (Rater # 1) had significant difference between her pre and post-test *Total Scores*, further analysis conducted on the difference between the pre and post-test *Total Scores* of students revealed that while there was absolute agreement between the 25 % of the pre and post-test scorings, 75 % of the *Total Scores* ranked lower or higher in the post-test. To be more precise, out of 168 scorings assigned by 15 raters, in terms of the Total Scores, 68 lower scores and 58 higher scores were observed while 42 of the scores did not change. Moreover, some of the differences between the pre and post-test *Total Scores* were more than one point which can be considered as a big difference because (a) the highest score that can be assigned was 20 points, (b) the raters' were informed that they were required to assess the students as if they were assessing an actual proficiency exam oral interview, (c) the proficiency exams are also considered as a high stake exam for the students in the institution where the study was conducted because they are used to decide whether the students are proficient enough to pursue their major studies, and most importantly (d) if the students fail, they are required to take the intensive English preparatory class one more year. For all these reasons, even a one point difference becomes important given that achieving high reliability also becomes very important especially when the decision or the result of a test is very important for the candidates (Hughes, 2003). Moreover, as discussed by Myford and Wolfe (2000), it should be noted that although one point may not seem like or be considered as a large difference, it can have an important effect for the test takers whose scores are around borderline/pass score.

Despite the change in the 75 % of the assigned scores, basing the study only on the results of the quantitative data is not enough to say that the raters' in the present study were affected by their prior knowledge of students' proficiency levels and assigned scores accordingly. As mentioned before, the fact that the human raters do the scorings in performance assessment has been at the center of the discussion because it has been acknowledged that raters may yield to subjectivity which may affect the ratings in oral interviews (Caban, 2003), and the previous studies conducted on the assessment of oral performances revealed that there are various factors that affect the scores such as the candidates, the rubric, the test itself, and the raters (Bachman, 1990; McNamara, 1997). Moreover, the results of the previous studies revealed that while some changes in the scores, in other words, the error of measurement, can be considered as systematic, some are random (Upshur & Turner, 1999). As a result, similar to the findings of the studies conducted on rater effects, differences were observed in the raters' pre and post-test assessment in the present study, but further analysis on the raters' verbal reports was conducted in relation to the scores they assigned in order to analyze whether the error of measurement was systematic rather than random. In other words, whether the raters' knowledge of the students' proficiency levels had an effect on their scorings was investigated by analyzing what the raters reported while assigning scores.

When each rater's verbal reports provided during the pre and post-test were analyzed, it was observed that 11 raters referred to the proficiency levels of the students during scoring for their performances while no reference to the levels of the students were observed in four raters' verbal reports although there were also lower and/or higher scores in their scorings. Figure 17 presents the results about whether the raters had significant difference in their scorings and/or referred to the proficiency levels of the students.

Rater No	Significant Difference	Reference to the levels
1[69]	**YES**	**YES**
2	**NO**	**YES**
3	**YES**	**YES**
4	**YES**	**YES**
5	**NO**	**NO**
6	**YES**	**YES**
7	**NO**	**YES**
8	**YES**	**YES**
9	**NO**	**YES**
10	**NO**	**YES**
11	**YES**	**NO**
12	**NO**	**YES**
13	**NO**	**NO**
14	**YES**	**YES**
15	**YES**	**NO**

Figure 17. The existence of significant difference in raters' scorings and/or reference to the proficiency levels in their verbal reports.

Since no reference to the levels were found in the four raters' verbal reports, the results are inconclusive for these raters due to the fact that the measurement error can be random or reference to the levels were not observed in their reports due to the "incompleteness due to synchronization problems" (Van Someren et al., 1994, p. 33). In other words, the variable in the post-test, raters' knowledge of students' proficiency levels did not affect their scorings or these raters may not have verbalized what they thought exactly, so there may be some missing data in their reports due to the difference between the pace they think and they speak (Van Someren et. al, 1994). However, it can be concluded that the 11 raters who referred to the proficiency levels of the students assigned higher or lower post-test *Total*

[69] Although lower and/or higher scorings were frequently observed in the post-test *Total Scores* assigned by each rater, Rater # 1 is the only rater who showed statistically significant difference between her pre and post-test *Total Scores*.

Scores to individual students when the information of the students' proficiency levels was provided in the post-test. The verbal reports of these raters indicated that all these raters used statements such as *"Good vocabulary such as get stressed, cough, it was good considering she is a D Level student"* (Rater # 6, Student # 12) which may suggest that the raters assigned scores to the performances of the students considering the proficiency levels of them and judging what each level of student could achieve in *Fluency and Pronunciation, Vocabulary*, and other aspects of the rubric. Some raters also assessed the success of the performances by referring to what each level of student could achieve in terms of the content of the curriculum they were taught during the year as seen in the reports of Rater # 7 provided for Student # 2: "*She could tell her ideas only by using adjectives"* (the pre-test *Vocabulary* score was 2 points and the *Total Score* was 10 points) and "*The student had adequate vocabulary range for this level of student"* (the pre-test *Vocabulary* score was 4 points and the *Total Score* was 15 points).

When the pre and post-test *Total Scores* assigned by these 11 raters were investigated in terms of their degree of severity/leniency towards lower and higher proficiency level students, it was observed that the raters behaved differently when the information about students' proficiency levels was provided in the post-test. While Rater # 2, Rater # 8, Rater # 9, and Rater # 12 assigned lower *Total Scores* for D level students, Rater # 1 was more severe in her scorings. For C levels, while Rater # 4, Rater # 8, and Rater # 14 were more severe, Rater # 1, Rater # 3, and Rater # 6, assigned more favorable scores in the post-test. B level students received harsher scorings from Rater # 3, Rater # 4, Rater # 9, Rater # 12, Rater # 14 while Rater # 1, and Rater # 13 were more lenient towards B level students.

There may be several reasons for why each rater perceived the performances of the students differently in the pre and post-test and so differed in their interpretations of the students' performances and degree of severity by assigning lower or higher post-test scores. The types of rater effects on scores described in the literature as halo effect, central tendency, restriction of range, and severity/leniency (Saal et al., 1980) may be helpful in explaining the rater variance observed in the findings of the present study. First, it can be the result of halo effect. In other words, with the knowledge of the students' proficiency levels, the raters may have assigned scores with "a global impression of each examinee" (Borman as cited in Saal et al., 1980, p. 415) rather than distinguishing different levels of performances in different aspects such as *Vocabulary, Grammatical Range and Accuracy*. For example, for Student # 10, Rater # 9 assigned scores to the components of the rubric from both lower and higher bends and a *Total Score* of 13 stating *"The student was less successful compared to her partner, but in pair work, it was obvious that this was a pair work, they asked questions to each other."* However, in the post-test, he mostly assigned scores from lower bends adding up to 10 points as a *Total Score* commenting *"Her partner continued the conversation although he was a D level student, but this student was passive although she was a B level student, and she performed less successfully than her partner."* As seen in the example, when the information about the student's proficiency level was available in the post-test, the rater assigned lower scores only in the *Task Completion* and *Comprehension* components which were the only two aspects that the student received the highest scores 3 and 4, respectively, in the pre-test. As a result, with the higher expectations from a B level student, the rater might have assigned lower scores for *Task Completion* and *Comprehension* in the post-test considering her level and poor performance in the other aspects. In short, the students' poor or better performance in one aspect may have affected the judgment of the raters if they considered the proficiency levels of the students while assigning scores.

Second, “raters’ reluctance to make extreme judgments” about the students which is called central tendency (Saal et al., 1980, p. 417), and similarly, raters’ overusing certain categories in each category of the rubric which is called the restriction of range (Myford & Wolfe, 2003) may have an effect on the differences in their scorings. The variance in the scores can be the effect of raters’ considering the levels of the students and what scores other raters would assign for these students. In other words, although they did not report such considerations verbally, novice raters or raters who did not want to stand out may have yielded to the effect of central tendency (Saal et al., 1980, p. 417) and the restriction of range. For example, for Student # 10, Rater # 1 assigned 5 points as a *Total Score* in the pre-test which was the lowest point that could be assigned and commented *“The student’s performance was very bad, she could not speak at all.”* However, in the post-test, the rater assigned 13 points as a *Total Score* stating *“Although the student is a B level student, she could not speak and could not do the task.”* As seen in the example, the rater assigned the lowest score in the pre-test, but her score in the post-test was around midpoint which might be the effect of rater’s considering that the student might receive higher scorings from other raters because she is a B level student - the highest level in the institution where the study was conducted. As a result, since the raters were aware that the data provided from all the raters would be analyzed, there is a possibility that, even if they used the lowest or the highest bends in the pre-test, they assigned scores around midpoint in the post-test in order not to differ from the other raters’ in terms of their degree of severity/leniency.

The most common type of rater effects on scores is severity/leniency. When the scores assigned for individual students were analyzed, it was observed that the raters exercised some degree of severity/leniency when rating students although there was no pattern in their assigning lower or higher scores for a specific level which would show evidence of rater bias towards a particular group of candidates. Central tendency which was discussed in the previous paragraph may also be helpful in understanding the reasons for severe ratings assigned for lower level students and favorable scorings for higher level students' *Total Scores*. In other words, raters may have avoided assigning scores from the highest bends for lower levels and scores from the lowest bends for higher levels considering the proficiency levels of the students and what scores other raters might assign for these students. One of the explanations for the changes in the raters' pre and post-test scorings which was mentioned earlier may also be the reason for the variations in the degree of severity/leniency the raters practiced. It was observed that the raters assessed the success of the performances of the students in terms of the content of the curriculum they were taught during the year. Although all the students took the same proficiency exam, the content of the instruction provided in the institution differs for lower levels and higher levels. This may have affected the raters' judgments in terms of their appreciation of the lower-level students' efforts and disgracing the higher level students' lack of enthusiasm and participation due to the higher expectations from a higher level student. A previous example provided is an indication of favorable ratings for lower levels. For a C level student, Student # 2, Rater # 7 assigned 2 points for the pre-test *Vocabulary* saying "*She could tell her ideas only by using adjectives." and* 10 points as the *Total Score* reporting "*The student was nervous in the first task, so she could not speak much. In the second task, although she had errors in her sentences, she told her ideas."* However, a favorable judgment was observed in the post-test. The rater assigned 4 points for *Vocabulary* pointing out *"The student had adequate vocabulary range for this level of student",* and 15 points as the *Total Score* commenting *"The student tried, but her sentence constructions were problematic, so even if she had a better performing partner, I don't think she can express herself well, still she completed her tasks."* However, for Student # 2, another rater, Rater # 2 showed severity in her scorings since she considered that C level is a higher level than her partner's D level. In the pre-test, she assigned 18 points as a *Total Score* for Student # 2 reporting *"She was excited in the first task, but she could formulate some sentences. It could be better. The second task was very successful, she asked all the questions and used all the necessary words. She initiated the conversation and it was very effective."* Yet, a great degree of severity was observed in her post-test scorings and verbal reports when the information about the students' levels was provided. The rater considered the level of the student as a higher level compared to her partner, and she assigned 8 points as a Total Score commenting *"Although she was a C level, the student was very excited. She had limited vocabulary range and grammar errors even in simple sentences which obscured the meaning. She had lots of pauses, so she had problems in fluency."* As a result, raters cannot be directly compared in terms of the degree of severity they exercise when scoring, but the knowledge of the students' proficiency levels could have affected each rater's degree of leniency or severity to some extent.

As seen in the examples above, another possible rationale behind these results is the fact that the students were paired randomly without considering their proficiency levels, and although very few took the exam with a same proficiency level student, most of the pairs included students with different proficiency levels. The analysis of verbal reports also revealed that some raters compared the performances of the two candidates taking the exam together as pairs by referring to their levels and assigning scores accordingly. The results indicated that this may have an effect on the changes of the scores because some raters assigned scores in the post-test considering the performances and the proficiency levels of the candidates and their partners as seen in many cases and in the example from Rater # 8's scorings and verbal reports for a pair, Student # 11, a D level student and Student # 12, a B level student. In the pre-test, Rater # 8 assigned 16 points as a *Total Score* for Student # 11 and 18 points for Student # 12 commenting *"Student 11's partner was a little better than him, especially in the first task, so she got 2 points higher than him."* However, when the information about students' proficiency levels was provided in the post-test, Rater # 8 differed in his scorings and verbal reports. He assigned 18 points for Student # 11 and 16 points for Student # 12 stating *"Student # 12's partner was more fluent, enthusiastic. Generally female students are more excited. They were successful. We should also consider that this student is a D level student, and the other one is a B level student."* As a result, even though the proficiency level might not be a variable on its own, when combined with pairs from different levels, it does seem like it makes a difference.

In conclusion, the findings of the present study contribute to the previous studies conducted on rater effects and have also verified that raters may sometimes be affected by the factors other than the actual performance of the candidates (e.g., Chalhoub-Deville, 1995; Chalhoub-Deville & Wigglesworth, 2005; Lumley & McNamara, 1995; Myford & Wolfe, 2000; Thompson, 1991; Winke & Gass, 2012). Whether random or systematic, it is no surprise that measurement error was observed in this study because oral performance assessment is such a complex procedure in which there are several influential factors that may cause disagreement within and/or among the raters’ judgments (McNamara, 1996). In light of the findings of the present study and the existing literature, it can be argued that the raters’ prior knowledge of students’ proficiency levels could be an important factor that may cloud raters’ judgments and affect their scoring behaviors during oral interview assessment which jeopardizes the assurance of the two important qualities of a good test: reliability and fairness (Bachman & Palmer, 1996; Kunnan, 2000).

Implications for Testing and Pedagogy

The findings of the present study point out important implications for testing and pedagogy that can inform the institutions, teachers and raters that assess the oral performances of the students in oral interviews. Because teaching and testing are two inseparable aspects of education (Rudman, 1989), the utmost care should also be given on testing due to the fact that no matter how perfect instruction is provided to the students, if there are some factors that affect the results of the tests other than the actual performance of the students, the goals of achieving success can never be reached. Regarding the fact that several factors affect the assessment of oral interviews, and the existence of human raters in oral interviews is one of the challenging factors that can change a score assigned to a test performance (Hardacre & Carris, 2010), the results of this study revealed that the knowledge of students’ proficiency levels is one of the factors that may jeopardize the results of the assessment, the reliability of the institutions where the assessment is done, and the academic and even personal lives of the students. For this reason, considering the detrimental effects of the raters’ prior knowledge of students’ proficiency levels on raters’ scorings, some recommendations can be made for the institutions to ensure that the effects of the construct irrelevant factors on the scorings are minimized.

First of all, the commonly accepted suggestions to increase rater reliability and fairness such as rater training (e.g., Brown, 2004; Hughes, 2003; Lumley & McNamara, 1995; Myford & Wolfe, 2000), using multiple raters as assessors (Council of Europe, 2001; Hughes, 2003), using a validated appropriate rubric (Hughes, 2003), introducing the rubric to the raters in detail (Bachman, 1990), and providing the same explicit and thorough instruction for all the raters on how to assess the students' performances in terms of what to expect and what to focus should be noted. In light of the assessment behaviors of the raters both during the norming sessions and in the exams, first, rater profiles should be created in order to investigate whether the raters are severe or lenient assessors by nature and to inform the raters about their scoring performances. Then, since using multiple raters as assessors is highly suggested in the literature (Council of Europe, 2001; Hughes, 2003), raters should be paired according to their profiles created. In terms of fairness, it is better to match a severe rater with a lenient one instead of having two severe or lenient assessors for the same test-taker. Since paired interviews are widely used, the candidates may be asked to interact with a professional interlocutor rather than with a fellow candidate, but the advantages and disadvantages of using this format should be considered thoroughly (Hughes, 2003). More importantly, any information about the candidates that can lead to subjective scorings should not be provided to the raters either by the candidates or the institutions (Hughes, 2003), and the raters should only base their judgments on the performances of the test takers and the rubric they use (Council of Europe, 2001).

Limitations of the Study

There are several limitations to this quasi-experimental study suggesting that the findings should be treated with caution. Initially, the focus of the study, rater effects, is the major limitation of the study since it has been acknowledged that the existence of human raters is the major reason of subjective scoring. Moreover, although great care was taken in order to create similar assessment conditions, there is a chance that the raters may not have behaved in the way they usually assign scores since they were aware that their scorings and verbal reports would be analyzed by the researcher. Also, the raters may have had a tendency to pay extra attention to the scores they assigned in the pre-test since they were informed that there would be another scoring session. However, it should be noted that to minimize the possible recall effect, (a) there was at least a five week interval between the pre and post-test treatment, (b) the raters were not informed about the actual purpose of the study, and (c) they were not told that they were going to assign scores for the same students in the post-test.

Additionally, basing the study on only one form of qualitative data gathered from the raters' verbal reports provided during think-aloud protocols can be another limitation. The raters may not have verbalized what they thought exactly since the process of thinking and speaking are not at the same pace in human cognition, and the raters were fully aware that the data they provided would be analyzed. For this reason, the lack of follow-up interviews with the raters should also be noted as a limitation since they could have been helpful to gather more information about what raters thought while assigning scores and why they had variations in their scorings and verbal reports.

Furthermore, sampling is another limitation of the study. First, the study was conducted in only one setting. The raters who were the participants in this study were all working at the same institution. It is possible that the findings may differ if the study was conducted with raters from different institutions. Second, although all the raters have had teaching and assessment experience of oral ability for at least one year, they did not receive any professional training for oral assessment and they were not certified raters.

The limited number of the available video-recordings of previous years' proficiency exam oral interview samples was another limitation to this study. Due to the poor audio quality of the recordings, the researcher did not have the opportunity to have oral interview samples which included equal number of students from three proficiency levels. Furthermore, since some raters were familiar with one or more students, the data gathered from those raters provided for the students they were familiar with could not be used in order to eliminate the effect of familiarity which is considered as a variable that may affect raters' scorings.

Last but not least, even though several attempts have been made, the study was also limited in its ability to control all the construct-irrelevant variables that might influence the assessment behaviors of the raters. To illustrate, although the researcher tried to choose the available video samples with the best quality, the audio quality of the videos can be one of these factors. The recall effect might be another factor though five week interval has been allotted between the pre and post-test. The format of the oral interviews in this study may also be another influential variable. Since the candidates were assigned two tasks for which they performed as individual test-taker and paired candidates, both the employment of two tasks and the performances of the candidates' partners may have affected the scorings of the raters although there was no change in the tasks and pairing of the students in the post-test.

Suggestions for Further Research

On the basis of the findings and the limitations of the study, some suggestions can be made for further research. To begin with, the study was conducted in one setting, so this study can be replicated in another setting or with participants from different institutions and backgrounds to reach at more generalizable findings and to see whether the findings are resulted from the effect of the setting and the lack of rater training. The number of the raters who assign scores and the number of students whose performance are assessed can be increased. Secondly, the same topic could be explored with a longer interval between the pre and post-test in order to ensure that the recall effect is successfully controlled. Third, the same study can be conducted by including follow-up interviews in the data collection and analysis process in order to gain more insights for why raters assign what they assign. Moreover, the study can also be replicated with a change in the methodology by carrying out the study with a treatment and a control group. While the information about the students' proficiency levels can be provided to the treatment group in the post-test, no information can be given to the control group in order to analyze if there is a significant difference between their scorings. Furthermore, since the sources of rater effects in oral performance assessment are at exploratory stage, the effect of any construct-irrelevant factor that has not been studied before can also be a potential research topic for further studies. Finally, all the suggestions mentioned above can also be applied to the written performance assessment since rater effects are also explored for writing exams.

Conclusion

This quasi-experimental study, conducted with 15 raters investigated whether the raters' prior knowledge of the students' proficiency levels had any effect on their scorings. The findings revealed that when the information about the students' proficiency levels were provided to the raters, 75 % of the scorings changed in the post-test as lower or higher scores, and 11 raters, in their verbal reports, referred to the proficiency levels of the students while assigning scores in the post-test. The findings of the study are in accordance with the literature which suggests that the construct-irrelevant factors can influence the assessment of the raters and the scores of the test-takers in oral interviews (e.g., Brennan & Brennan, 1981; Carey et al., 2010; Chalhoub-Deville, 1995; Chalhoub-Deville & Wigglesworth, 2005; Derwing & Munro, 1997; Galloway, 1980; Gholami et al., 2011; Lumley & McNamara, 1995; Myford & Wolfe, 2000; O'Loughlin, 2002; O'Sullivan, 2000; Thompson, 1991; Winke & Gass, 2012; Winke et al., 2011).

Several factors that affect raters' scorings in oral interviews have been studied in the literature; however, to the knowledge of the researcher, no study has been conducted to investigate the effects of the raters' prior knowledge of the students' proficiency levels on their scoring behaviors during proficiency exams oral interviews. Therefore, this study might augment the literature by revealing another source of rater effects in oral interviews assessment. To conclude, it is hoped that the findings of the study and the pedagogical implications discussed in this chapter will help all the stakeholders gain insight into the importance of minimizing any external factor that may jeopardize the reliability and the fairness of the scorings assigned for the students.

References

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

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

Bachman, L. F. (2004). *Statistical analyses for language assessment*. New York: Cambridge University Press.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing, 12*(2), 238-257. doi: 10.1177/026553229501200206

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

Barnwell, D. (1989). 'Naive' native speakers and judgements of oral proficiency in Spanish. *Language Testing, 6*(2), 152-163. doi: 10.1177/026553228900600203

Boulet, J. R., Van Zanten, M., McKinley, D. W., & Gary, N. E. (2001). Evaluating the spoken English proficiency of graduates of foreign medical schools. *Medical Education, 35*(8), 767-773. doi: 10.1046/j.1365-2923.2001.00998.x

Brennan, E. M., & Brennan, J. S. (1981). Accent scaling and language attitudes: Reactions to Mexican American English speech. *Language and Speech, 24*(3), 207-221. doi: 10.1177/002383098102400301

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing, 26*(3), 341-366. doi: 10.1177/0265532209104666

Brown, A. (1995). The effect of rater variables in the development of an occupationspecific language performance test. *Language Testing, 12*(1), 1-15. doi: 10.1177/026553229501200101

Brown, G., & Yule, G. (1999). *Teaching the spoken language.* Cambridge: Cambridge University Press.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. New York: Longman.

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

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

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater’s familiarity with a candidate’s pronunciation affect the rating in oral proficiency interviews? *Language Testing, 28*(2), 201-219. doi: 10.1177/0265532210393704

Carrell, P. L. (1995). The effect of writers' personalities and raters' personalities on the holistic evaluation of writing. *Assessing Writing, 2*(2), 153-190. doi: 10.1016/1075-2935(95)90011-x

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*(1), 16-33. doi: 10.1177/026553229501200102

Chalhoub-Deville, M., & Wigglesworth, G. (2005). Rater judgment and English language speaking proficiency. *World Englishes, 24*(3), 383-391. doi: 10.1111/j.0083-2919.2005.00419.x

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.

Council of Europe. (2001) *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press. Retrieved from: http://www.coe.int/t/dg4/linguistic/source/framework_en.pdf

Crocker, L. M., & Algina, J. (1986). *Introduction to classical & modern test theory*. Orlando, FL: Holt, Rinehart, and Winston.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing, 26*(3), 367-396. doi: 10.1177/0265532209104667

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition, 19*(01), 1-16.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*(3), 197-221. doi: 10.1207/s15434311laq0203_2

Ellis, R. O. D., Johnson, K. E., & Papajohn, D. (2002). Concept mapping for rater training. *TESOL Quarterly, 36*(2), 219-233. doi: 10.2307/3588333

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215-251. doi: 10.1037/0033-295X.87.3.215

Farrokhi, F., & Esfandiari, R. (2011). A many-facet Rasch model to detect halo effect in three types of raters. *Theory and Practice in Language Studies, 1*(11), 1531-1540. doi:10.4304/tpls.1.11.1531-1540

Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning, 37*(3), 313-326. doi: 10.1111/j.1467-1770.1987.tb00573.x

Fulcher, G. (2003). *Testing second language speaking.* London: Longman/Pearson Education.

Fulcher, G., & Bamford, R. (1996). I didn't get the grade I need. Where's my solicitor? *System, 24*(4), 437-448. doi: http://dx.doi.org/10.1016/S0346-251X(96)00040-1

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment.* London, New York: Routledge.

Galloway, V. B. (1980). Perceptions of the communicative efforts of American students of Spanish. *The Modern Language Journal, 64*(4), 428-433. doi: 10.1111/j.1540-4781.1980.tb05218.x

Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning, 34*(1), 65-87. doi: 10.1111/j.1467-1770.1984.tb00996.x

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.

Hadden, B. L. (1991). Teacher and nonteacher perceptions of second-language communication. *Language Learning, 41*(1), 1-20. doi: 10.1111/j.1467-1770.1991.tb00674.x

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27. doi: 10.1111/j.1745-3992.2004.tb00149.x

Hardacre, B., Carris, L. (2010). The UCLA test of oral proficiency: A model for assessing and addressing English proficiency of international teaching assistants. In Avineri, N., Londe, Z., Hardacre, B., Carris, L., So, Y., &

Majidpour, M. (Eds). Language assessment as a system: Best practices, stakeholders, models, and testimonials. *Issues in Applied Linguistics, 18*(2). Retrieved from: http://escholarship.org/uc/item/9c20c0wn

Hill, K. (1996). Who should be the judge? The use of non-native speakers as raters on a test of English as an international language. *Melbourne Papers in Language Testing, 5*(2), 29-50.

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

Iwashita, N. (1997). *The validity of the paired interview format in oral performance testing*. Paper presented at 19th Annual Language Testing Research Colloquium. Florida, USA.

Jacobs, G. M., & Farrell, T. S. C. (2003). Understanding and implementing the CLT (Communicative Language Teaching) paradigm. *RELC Journal*, *34*(1), 5-30. doi: 10.1177/003368820303400102

Joe, J. N. (2008). *Using verbal reports to explore rater perceptual processes in scoring: An application to oral communication assessment* (Unpublished doctoral dissertation). James Madison University, Assessment and Measurement, Harrisonburg, VA, USA.

Joe, J. N., Harmes, J. C., & Hickerson, C. A. (2011). Using verbal reports to explore rater perceptual processes in scoring: A mixed methods application to oral communication assessment. *Assessment in Education: Principles, Policy & Practice, 18*(3), 239-258. doi: 10.1080/0969594x.2011.577408

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education. Washington, DC.* Retrieved January 08, 2013 from http://www.apa.org/science/programs/testing/fair-code.aspx#c

Joughin, G. (1998). Dimensions of oral assessment. *Assessment & Evaluation in Higher Education, 23*(4), 367-378. doi: 10.1080/0260293980230404

Kang, O. (2012). Impact of rater characteristics and prosodic features of speaker accentedness on ratings of international teaching assistants' oral performance. *Language Assessment Quarterly, 9*(3), 249-269. doi: 10.1080/15434303.2011.642631

Kim, Y.-H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing, 26*(2), 187-217. doi: 10.1177/0265532208101010

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 1-14). Cambridge, UK: Cambridge University Press.

Kuusela, H., & Paul, P. (2000). A comparison of concurrent and retrospective verbal protocol analysis. *American Journal of Psychology, 113*(3), 387-404.

Lazaraton, A., & Riggenbach, H. (1990). Oral skills testing: A rhetorical task approach. *Issues in Applied Linguistics*, 1(2). Retrieved from: http://www.escholarship.org/uc/item/77h2z2xh

Linacre, J. M. (1994). *Many-facet Rasch measurement*. Chicago, IL: Mesa.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54-71. doi: 10.1177/026553229501200104

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

MacIntyre, P. D., Noels, K. A., & Clément, R. (1997). Biases in self-satings of second language proficiency: The role of ranguage anxiety. *Language*

Learning, 47(2), 265-287. doi: 10.1111/0023-8333.81997008

McNamara, T. F. (1990). *Assessing the second language proficiency of health professionals* (Unpublished doctoral dissertation). The University of Melbourne, Department of Linguistics and Language Studies, Victoria, Australia.

McNamara, T. F. (1996). *Measuring second language performance.* New York: Longman.

McNamara, T. F. (1997). ‘Interaction’ in second language performance assessment: Whose performance? *Applied Linguistics, 18*(4), 446-466. doi: 10.1093/applin/18.4.446

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

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA & Oxford: Blackwell.

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition, 28*(01), 111-131. doi: doi:10.1017/S0272263106060049

Myford, C. M., & Wolfe, E. W. (2000). *Monitoring sources of variability within the Test of Spoken English Assessment System. TOEFL Research Report 65*. Princeton, NJ: TOEFL Research Program, Educational Testing Service.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386–422.

O’Loughlin, K. J. (1997). *Direct and semi-direct tests of spoken language* (Unpublished doctoral dissertation). University of Melbourne, Melbourne, Australia.

O’Loughlin, K. J. (2001). *The equivalence of direct and semi-direct speaking tests*. Cambridge, UK: Cambridge University Press.

O’Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing, 19*(2), 169-192. doi: 10.1191/0265532202lt226oa

Orr, M. (2002). The FCE speaking test: using rater reports to help interpret test scores. *System, 30*(2), 143-154. doi: http://dx.doi.org/10.1016/S0346-251X(02)00002-7

O'Sullivan, B. (2000). Exploring gender and oral proficiency interview performance. *System, 28*(3), 373-386. doi: http://dx.doi.org/10.1016/S0346-251X(00)00018-X

O’Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing, 19*(3), 277-295. doi: 10.1191/0265532202lt205oa

Oztekin, E. (2011). *A comparison of computer assisted and face-to-face speaking assessment: Performance, perceptions, anxiety, and computer attitudes* (Unpublished master’s thesis). Bilkent University, MATEFL Program, Ankara, Türkiye.

Powers, D. E., Schedl, M. A., Leung, S. W., & Butler, F. A. (1999). Validating the revised Test of Spoken English against a criterion of communicative success. *Language Testing, 16*(4), 399-425. doi: 10.1177/026553229901600401

Pula, J.J., & Huot, B.A. (1993). A model of background influences on holistic raters. In M.M. Williamson, & B.A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 237-265). Cresskill, NJ: Hampton Press.

Qian, D. D. (2009). Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers. *Language Assessment Quarterly, 6*(2), 113-125. doi: 10.1080/15434300902800059

Richards, J. C., & Schmidt, R. W. (2010). *Longman dictionary of language teaching and applied linguistics* (4th Ed.). Harlow: Longman.

Rudman, H. C. (1989). Integrating testing with teaching. *Practical Assessment, Research & Evaluation*, *1*(6). Retrieved December 4, 2012 from http://PAREonline.net/getvn.asp?v=1&n=6

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413-428. doi: 10.1037/0033-2909.88.2.413

Stansfield, C. W., & Kenyon, D. M. (1992). The development and validation of a simulated oral proficiency interview. *The Modern Language Journal, 76*(2), 129-141. doi: 10.2307/329767

Stoynoff, S. (2012). Research agenda: Priorities for future research in second language assessment. *Language Teaching, 45*(02), 234-249. doi:10.1017/S026144481100053X

Sunderland, J. (2000). Issues of language and gender in second and foreign language education. *Language Teaching, 33*(04), 203-223. doi:10.1017/S0261444800015688

Tedick, D. J. (1990). ESL writing assessment: Subject-matter knowledge and its impact on performance. *English for Specific Purposes, 9*(2), 123-143. doi: http://dx.doi.org/10.1016/0889-4906(90)90003-U

Thompson, I. (1991). Foreign accents revisited: The English pronunciation of Russian immigrants. *Language Learning, 41*(2), 177-204. doi: 10.1111/j.1467-1770.1991.tb00683.x

Upshur, J. A., & Turner, C. E. (1999). Systematic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing, 16*(1), 82-111. doi: 10.1177/026553229901600105

Ur, P. (1999). *A course in language teaching: Trainee book*. Cambridge: Cambridge University Press.

Van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The think aloud method: A practical guide to modeling cognitive processes*. London: Academic Press.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197-223. doi: 10.1177/026553229401100206

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6*(2), 145-178. doi: http://dx.doi.org/10.1016/S1075-2935(00)00010-6

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

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing, 10*(3), 305-319. doi: 10.1177/026553229301000306

Wigglesworth, G. (2001). Influences on performance in task-based oral assessments. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks second language learning, teaching and testing* (pp. 186–209). Essex, UK: Longman

Wiliam, D. (2008). Quality in assessment. In S. Swaffield (Ed.), *Unlocking assessment: Understanding for reflection and application.* (pp. 123-137). Abingdon: Routledge.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah, NJ: Lawrence Erlbaum.

Winke, P., & Gass, S. (2012). The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation. *TESOL Quarterly*, n/a-n/a. doi: 10.1002/tesq.73

Winke, P., Gass, S., & Myford, C. (2011). The relationship between raters' prior language study and the evaluation of foreign language speech samples. *TOEFL iBT® Research Report*. Princeton, NJ: Educational Testing Services. Retrieved from: http://www.ets.org/Media/Research/pdf/RR-11-30.pdf.

APPENDICES

Appendix 1 – Informed Consent Form

INFORMED CONSENT FORM

Dear Colleague;

I am Fatma TANRIVERDİ KÖKSAL, one of the instructors of English at Bülent Ecevit University Foreign Languages Compulsory Preparatory Program. I have been doing MA degree in the department of Teaching English as a Foreign Language at İhsan Doğramacı Bilkent University. The purpose of my thesis subject is to investigate the decision making process the raters go through while assigning scores during oral interviews in proficiency exams.

In this study, the information about what raters think and perceive during assigning scores will be acquired through participants' scorings and think-aloud protocols. You are required to participate in two scoring sessions and verbalize what you think during assigning scores for six pairs of students in pre-recorded videos during 2011-2012 proficiency exam. The information about your identification will be kept confidential and will not be published in any reports at the end of the research.

Your participation will contribute to the study to a great extent. If you accept taking part in this study, please fill in the related blanks at the bottom of this page and sign.

Fatma TANRIVERDİ-KÖKSAL
Supervisor: Dr. Deniz ORTAÇTEPE
MA TEFL, İhsan Doğramacı Bilkent University / ANKARA

I have read the information in this form, and I accept participating in the study. I agree to the think-aloud protocols being video recorded.

Name and Surname:…….…………………….……

(Your signature below means that you voluntarily agree to participate in this thesis study.)

Signature:……………………………… Date: 16/01/2013

Appendix 2 –Demographic Information Questionnaire

Dear Colleague;

This questionnaire was designed to get some background information (e.g., educational, professional) about the raters participating in this thesis study. The answers you give will be analyzed taking your privacy into account. **Please do not leave any of the questions unanswered.**

1) Age: …………………

2) Gender: **a)** Male **b)** Female

3) Graduated BA program:

a) English Language Teaching
b) English Language and Literature / American Culture and Literature
c) Translation and Interpretation
d) Linguistics
e) Other ………………………………

4) MA degree: a) No **b)** Yes, Continuing **c)** Yes, Completed

<u>If yes, please specify your field:</u>

a) ELT
b) English Language and Literature / American Culture and Literature
c) Educational Sciences
d) Other: ……………………………….

5) PhD: **a)** No **b)** Yes, Continuing **c)** Yes, Completed

<u>If yes, please specify your field:</u>

a) ELT
b) English Language and Literature / American Culture and Literature
c) Educational Sciences
d) Other: ……………………………

6) Experience in teaching: .... years

7) How long have you been working at this institution?: .... years

8) How long have you been administering <u>proficiency exam</u> speaking tests as a rater?
.... years

Thank you for your participation.

MA TEFL student Fatma TANRIVERDİ KÖKSAL
fatmatanriverdi@gmail.com
Supervisor: Dr. Deniz ORTAÇTEPE

Appendix 3: Final Examination Speaking Rubric

Component	Fluency & Pronunciation	Vocabulary	Grammatical Range & Accuracy	Task Completion	Comprehension
4	Speaks smoothly, with little hesitation that does not interfere with communication. Pronunciation and intonation are almost always very clear/accurate.	Uses of vocabulary & conversational expressions accurate and appropriate.	Makes few (if any) noticeable errors of grammar or word order.	Topics dealt with comprehensively & relevantly with appropriate details.	Student appears to understand everything said; easy for the listener to understand student's intention and general meaning.
3	Speaks with some hesitation, but it does not usually interfere with communication. Pronunciation and intonation are usually clear / accurate with a few problem areas.	Appropriate terms used, but student must rephrase ideas due to lexical inadequacies.	Some errors of grammar & / word order, but meaning not obscured.	Topics dealt with comprehensively with limited details.	Student understands most everything said, yet repetition & clarification necessary.
2	Noticeable hesitations which sometimes disturb listener or prevent communication. Mispronunciations are frequent.	Communication limited from inadequate & inappropriate vocabulary.	Frequent errors of grammar and / or word order which obscure meaning.	Moderate success with topics; some details; some irrelevant data/ideas.	Student has difficulty in understanding what is said & requires frequent repetition.
1	Fragmentary and disconnected speech results in disrupted communication. Pronunciation and intonation errors sometimes make it difficult to understand the student.	Frequent misuse of words & very limited vocabulary and expressions.	Many errors, even in basic structures.	Limited success with topics; some details; includes irrelevant data/ideas.	Student has great difficulty in understanding what is said despite frequent repetitions.

Appendix 4 : Pre-Test Grading Sheet

RATER ________________ **Date:**____________________

Pair Task: What's your roommate like? (Great / Terrible)

STUDENT A: (LEFT)

ID: CLD837	
Familiarity: TAUGHT / OTHER / NO	

Task 1:	Love, Dating, & Marriage
Task 2:	Roommate?

GRADE for Student A	**Fluency & Pronunciation**		**Vocabulary**		**Grammatical Range & Accuracy**		**Task Completion**		**Comprehension**		**TOTAL:**	
		4		**4**		**4**		**4**		**4**		**20**

STUDENT B: (RIGHT

ID: VTM382	
Familiarity: TAUGHT / OTHER / NO	

Task 1:	Advertising
Task 2:	Roommate?

GRADE for Student B	**Fluency & Pronunciation**		**Vocabulary**		**Grammatical Range & Accuracy**		**Task Completion**		**Comprehension**		**TOTAL:**	
		4		**4**		**4**		**4**		**4**		**20**

NOTES:

STUDENT A: (LEFT)

STUDENT B: (RIGHT)

Appendix 5 : Post-Test Grading Sheet

RATER __________________ **Date:______________________**

Pair Task: What's your roommate like? (Great / Terrible)

STUDENT A: (LEFT)

ID: CLD837 **LEVEL: D**
Familiarity: TAUGHT / OTHER / NO

Task 1: Love, Dating, & Marriage
Task 2: Roommate?

GRADE for Student A	**Fluency & Pronunciation**		**Vocabulary**		**Grammatical Range & Accuracy**		**Task Completion**		**Comprehension**		**TOTAL:**	
		4		**4**		**4**		**4**		**4**		**20**

STUDENT B: (RIGHT

ID: VTM382 **LEVEL: C**
Familiarity: TAUGHT / OTHER / NO

Task 1: Advertising
Task 2: Roommate?

GRADE for Student B	**Fluency & Pronunciation**		**Vocabulary**		**Grammatical Range & Accuracy**		**Task Completion**		**Comprehension**		**TOTAL:**	
		4		**4**		**4**		**4**		**4**		**20**

NOTES:

STUDENT A: (LEFT)

STUDENT B: (RIGHT)

Appendix 6: Rater’s Scorings and Verbal Reports During the Pre and Post-Test

RATER X[70]										
Student No	**Level**	**Level of Pair**	**Pre & Post Test**	**Fluency & Pronunciation**	**Vocabulary**	**Grammatical Range & Accuracy**	**Task Completion**	**Comprehension**	**Total Score**	**Final Comments**
1 CLD837 FEMALE	**D**	**C**	Pre-test scores & comments							
			Post-test scores & comments							
2 VTM382 FEMALE	**C**	**D**	Pre-test scores & comments							
			Post-test scores & comments							
3 DZK178 MALE	**B**	**B**	Pre-test scores & comments							
			Post-test scores & comments							
4 WTL382 FEMALE	**B**	**B**	Pre-test scores & comments							
			Post-test scores & comments							

[70] D/C/B Levels: D is the lowest, B is the highest level. Yellow colored components of the rubric are the ones with significant difference. Red colored statements are references to the level of the students.

Student No	Level	Level of Pair	Pre & Post Test	Fluency & Pronunciation	Vocabulary	Grammatical Range & Accuracy	Task Completion	Comprehension	Total Score	Final Comments
5 **FTK139** **MALE**	**C**	**D**	Pre-test scores & comments							
			Post-test scores & comments							
6 **KMH532** **FEMALE**	**D**	**C**	Pre-test scores & comments							
			Post-test scores & comments							
7 **LSN792** **FEMALE**	**D**	**C**	Pre-test scores & comments							
			Post-test scores & comments							
8 **PFJ483** **MALE**	**C**	**D**	Pre-test scores & comments							
			Post-test scores & comments							
9 **TLS517** **MALE**	**D**	**B**	Pre-test scores & comments							
			Post-test scores & comments							

Student No	Level	Level of Pair	Pre & Post Test	Fluency & Pronunciation	Vocabulary	Grammatical Range & Accuracy	Task Completion	Comprehension	Total Score	Final Comments
10 HTN495 FEMALE	**B**	**D**	Pre-test scores & comments							
			Post-test scores & comments							
11 XRZ347 MALE	**B**	**D**	Pre-test scores & comments							
			Post-test scores & comments							
12 YTR790 FEMALE	**D**	**B**	Pre-test scores & comments							
			Post-test scores & comments							