

+



**FACTORS AFFECTING
TERTIARY LEVEL EFL LEARNERS'
SPEAKING PERFORMANCES**

PhD DISSERTATION

Meriç AKKAYA ÖNAL

Eskişehir 2025

**FACTORS AFFECTING TERTIARY LEVEL EFL LEARNERS' SPEAKING
PERFORMANCES**

Meri AKKAYA NAL

PhD DISSERTATION

Department of Foreign Language Education

PhD Program in English Language Teaching

Supervisor: Prof. Dr. zgr YILDIRIM

Anadolu University

Graduate School

June 2025

Eskişehir

FINAL APPROVAL FOR THESIS

This thesis titled “**Factors Affecting Tertiary Level EFL Learners’ Speaking Performances**” which has been prepared and submitted by Meriç AKKAYA ÖNAL in partial fulfillment of the requirements in “**Anadolu University Directive on Graduate Education and Examination**” for the PhD. in **Department of Foreign Language Education English Language Teaching** has been examined and approved on 02/06/20225.

<u>Committee Members</u>		<u>Signature</u>
Member (Supervisor) :	Prof. Dr. Özgür YILDIRIM	
Member :	Prof. Dr. Ali MERCİ	
Member :	Prof. Dr. Yusuf DEMİR	
Member :	Dr. Lecturer Hülya İPEK	
Member :	Dr. Lecturer Mustafa SARIOĞLU	

Prof. Dr. Nafiz Öncü CAN
Anadolu University
School of Graduate Studies Direct

JÜRİ VE ENSTİTÜ ONAYI

Meriç AKKAYA ÖNAL'ın “**Factors Affecting Tertiary Level EFL Learners’ Speaking Performances**” başlıklı tezi **02 Haziran 2025** tarihinde, aşağıdaki jüri tarafından Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin 37. Maddesi uyarınca **Yabancı Diller Eğitimi Anabilim Dalı İngilizce Eğitimi Programı’nda, Doktora** tezi olarak kabul edilmiştir.

		<u>Ünvanı Adı Soyadı</u>	<u>İmza</u>
Üye (Danışman)	:	Prof. Dr. Özgür YILDIRIM	
Üye	:	Prof. Dr. Ali MERÇ	
Üye	:	Prof. Dr. Yusuf DEMİR	
Üye	:	Dr. Öğr. Üy. Hülya İPEK	
Üye	:	Dr. Öğr. Üy. Mustafa SARIOĞLU	

Prof. Dr. Nafiz Öncü CAN
Anadolu Üniversitesi
Lisansüstü Eğitim Enstitüsü Müdürü

ÖZET

İNGİLİZCEYİ YABANCI DİL OLARAK ÖĞRENEN ÜNİVERSİTE ÖĞRENCİLERİNİN KONUŞMA PERFORMANSINI ETKİLEYEN FAKTÖRLER

Meriç AKKAYA ÖNAL

Yabancı Diller Eğitimi Ana Bilim Dalı

İngiliz Dili Eğitimi Programı

Anadolu Üniversitesi, Lisansüstü Eğitim Enstitüsü, Mayıs 2025

Danışman: Prof. Dr. Özgür YILDIRIM

Bu çalışma, yabancı dil eğitiminde iletişimsel yetkinliğe yapılan vurgunun artması bağlamında, öğrencilerin sözlü başarı testlerindeki konuşma performanslarını şekillendiren değişkenleri keşfetmeyi amaçlamaktadır. Bu olguya yanıt aramak için, derinlemesine araştırmaya olanak sağladığı düşünülen karma yöntem kullanılmıştır. Bu bağlamda, araştırma Türkiye'de bir devlet üniversitesinde gönüllülük esasına dayalı olarak çeşitli paydaşlarla gerçekleştirilmiştir. Bu katılımcılar arasında Yabancı Dil olarak İngilizce öğretmenleri, test tasarımcıları ve yöneticilerinin yanı sıra veri toplama sırasında beş dil yeterlilik seviyesindeki EFL öğrencileri de yer almaktadır.

Veriler yakınsak karma yöntem tasarımıyla toplandığından, hem nicel hem de nitel veriler paralel zamanlarda toplanmıştır. Nicel veriler için, açık uçlu bir soru ile 4'lü Likert ölçeği tüm paydaşlar tarafından yanıtlanmıştır. Nitel kısım için ise iki farklı araç kullanılmıştır. İlk olarak, tüm paydaşlarla yarı yapılandırılmış görüşmeler gerçekleştirilmiştir. İkinci olarak, İngilizceyi Yabancı Dil olarak öğrenen öğrenciler ve vize konuşma sınavındaki test değerlendiricileri ile video uyarımlı hatırlama görüşme protokolü aracılığıyla görüşülmüştür.

Bulgular, katılımcıların kaygı, gerginlik ve hata yapma korkusu gibi duygusal faktörleri sözlü başarıları üzerinde önemli bir engel olarak belirttiklerini ortaya koymuştur. Hazır bulunuşluk, yani sınıf içinde ve dışında pratik yapmak, neredeyse eşit derecede etkili bir faktör olarak belirtilmiş, bunu kelime bilgisi ve dilbilgisi yeterlilikleri gibi dilsel unsurlar izlemiştir. Buna ek olarak, sınırlı içerik üretimi, sınav ortamının fiziksel bağlamı, puanlayıcı davranışı ve puanlayıcı aşinalığı, görev soruları ve akran etkisinin öğrencilerin sözlü başarısı üzerinde önemli etkileri olduğu bulunmuştur.

Bu çalışma, başarı testleri sırasında öğrencilerin konuşma performansını etkileyen bu belirli faktörleri araştırarak, duygusal, dilsel ve bağlamsal değişkenlerin konuşma sınavı ortamlarında nasıl etkileşime girdiğinin daha derinlemesine anlaşılmasına katkıda bulunmaya çalışmaktadır.

Anahtar Kelimeler: Konuşma sınavları, Konuşma değerlendirmesi, İkinci dilde konuşma, İkinci dil konuşmada faktörler

ABSTRACT

FACTORS AFFECTING TERTIARY LEVEL EFL LEARNERS' SPEAKING PERFORMANCES

Meriç AKKAYA ÖNAL

Department of Foreign Language Education

PhD Program in English Language Teaching

Graduate School of Anadolu University, May 2025

Supervisor: Prof. Dr. Özgür YILDIRIM

In the context of growing emphasis on communicative competence in foreign language education, this study aims to explore the variables that shape learners' speaking performance in oral achievement tests. To seek answers to this phenomenon, a mixed-method methodology was used for in-depth research, and carried out at a public university in Türkiye with various stakeholders on voluntary basis including EFL instructors, test-designers, administrators, and EFL learners across five proficiency levels at the time of data collection.

The data were gathered in convergent mixed-methods design, where the quantitative and qualitative data were collected at parallel times. For the quantitative data, a 4-point Likert scale with an open-ended question was responded by all stakeholders. For the qualitative part, two different instruments were utilized. First, semi-structured interviews were carried out with all stakeholders. Second, EFL learners and testers in midterm speaking exam were interviewed through a video-stimulated recall interview protocol.

The findings yielded affective factors including anxiety, nervousness, and fear of making mistakes as a major hurdle on oral production. Readiness, namely practice within and beyond the classroom, was almost equally cited as an influential factor, followed by linguistic elements such as vocabulary and grammar competencies. Additionally, limited content production, physical context of the exam setting, rater behavior and familiarity, task questions, and peer effect were found to have significant effects on learners' oral achievement.

By exploring these specific factors, this study tries to contribute to a deeper understanding of how emotional, linguistic, and contextual variables interact in speaking exam settings.

Keywords: Speaking exams, Speaking assessment, L2 speaking, Factors affecting L2 speaking

24/06/2025

ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ

Bu tezin bana ait, özgün bir çalışma olduğunu; çalışmamın hazırlık, veri toplama, analiz ve bilgilerin sunumu olmak üzere tüm aşamalarında bilimsel etik ilke ve kurallara uygun davrandığımı; bu çalışma kapsamında elde edilen tüm veri ve bilgiler için kaynak gösterdiğimi ve bu kaynaklara kaynakçada yer verdiğimi; bu çalışmanın Anadolu Üniversitesi tarafından kullanılan “Bilimsel İntihal Tespit Programı”yla tarandığını ve hiçbir şekilde “intihal içermediğini” beyan ederim. Herhangi bir zamanda, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun saptanması durumunda, ortaya çıkacak tüm ahlaki ve hukuki sonuçları kabul ettiğimi bildiririm.

Meriç AKKAYA ÖNAL

STATEMENT OF COMPLIANCE WITH ETHICAL PRINCIPLES AND RULES

I hereby truthfully declare that this thesis/dissertation is an original work prepared by me; that I have behaved in accordance with the scientific ethical principles and rules throughout the stages of preparation, data collection, analysis and presentation of my work; that I have cited the sources of all the data and information that could be obtained within the scope of this study, and included these sources in the references section; and that this study has been scanned for plagiarism with “Scientific Plagiarism Detection Program” used by Anadolu University, and that “it does not have any plagiarism” whatsoever. I also declare that, if a case contrary to my declaration is detected in my work at any time, I hereby express my consent to all the ethical and legal consequences that are involved.

Meriç AKKAYA ÖNAL

ACKNOWLEDGEMENTS

Over the last few years of this challenging yet exciting journey, I would like to express my heartfelt gratitude to a great number of individuals who have helped and encouraged me a lot.

First and foremost, I would like to express my sincere gratitude to my supervisor, Prof. Dr. Özgür Yıldırım, for his continuous support, guidance, and encouragement throughout the research and writing process. His insightful feedback and expertise were invaluable at every stage of this thesis.

I would also like to thank my thesis advisory committee members, Prof. Dr. Ali Merç and Asst. Prof. Dr. Mustafa Sarioğlu, for their invaluable assistance and patience throughout the process.

I am deeply thankful to the faculty director, Prof. Dr. Ümit Deniz Turan, deputy directors, Inst. Mehmet Duranlıoğlu and Inst. Mehmet Gönen and the instructors and the students of the School of Foreign Languages at Anadolu University for providing a stimulating academic environment and the resources necessary to complete this work. This thesis would certainly not have been possible without them.

Special thanks to my colleague Revan Serpil, Sevil Ak Çimenler and Merve Nur Boldan who offered helpful discussions, feedback, and companionship along the way. Your support made the research journey more collaborative and enjoyable.

I am especially grateful to my husband, Barış Önal and my daughter Nehir Önal for their unwavering support, patience, and belief in me throughout this process. Without their encouragement and understanding, this thesis would not have been possible.

Thank you all.

24/06/2025

ÜRETKEN YAPAY ZEKÂ KULLANIM BEYANI

Bu tezi hazırlarken (örneğin ChatGPT, Gemini, DALL-E vb.) üretken yapay zekâ programlarından destek aldığımı/almadığımı beyan ederim. Tezin hazırlığı aşamasında üretken yapay zekâ programlarından (örneğin çeviri, bilimsel makale vb.) desteği aldım. Üretken yapay zekâ programlarından aldığım bilgilerin doğruluğunu kontrol ettiğimi bildiririm.

Herhangi bir zamanda, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun saptanması durumunda, ortaya çıkacak tüm ahlaki ve hukuki sonuçları kabul ettiğimi bildiririm.

Meriç AKKAYA ÖNAL

24/06/2025

DECLARATION OF GENERATIVE ARTIFICIAL INTELLIGENCE USAGE

I declare that I received support from generative artificial intelligence programs (for example ChatGPT, Gemini, DALL-E, etc.) while preparing this thesis/dissertation. I have received support from generative artificial intelligence programs (e.g. language translation, access to scientific articles, etc.) during the preparation of the thesis/dissertation. I declare that I have checked the accuracy of the information I received from generative artificial intelligence programs.

I also declare that, if a case contrary to my declaration is detected in my work at any time, I hereby express my consent to all the ethical and legal consequences that are involved.

Meriç AKKAYA ÖNAL

CONTENT PAGE

	<u>Page</u>
TITLE PAGE	
FINAL APPROVAL FOR THESIS	
ÖZET	iii
ABSTRACT.....	iv
ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ	v
STATEMENT OF COMPLIANCE WITH ETHICAL PRINCIPLES AND RULES	vi
ACKNOWLEDGEMENTS	vii
ÜRETKEN YAPAY ZEKÂ KULLANIM BEYANI	viii
DECLARATION OF GENERATIVE ARTIFICIAL INTELLIGENCE USAGE	ix
CONTENT	x
LIST OF TABLES	xv
LIST OF FIGURES	xvii
LIST OF ABBREVIATIONS	xviii
1 INTRODUCTION	1
1.1 Background to the study	1
1.2 Statement of the problem	6
1.3 Significance of the study	10
1.4 Purpose of the study	11
1.5 Definitions of key terms	12
2 LITERATURE REVIEW	14
2.1 What is a test?	16
2.2 Reasons for testing	16
2.3 Testing and assessment	17
2.4 Types of assessment	18
2.5 Kinds of language tests	20
2.6 Kinds of testing	22
2.7 Principles of language testing	24

2.7.1 Practicality	25
2.7.2 Reliability	26
2.7.3 Validity	28
2.7.4 Authenticity	32
2.7.5 Washback	33
2.8 Characteristics of test tasks	34
2.8.1 Characteristics of the setting	37
2.8.2 Characteristics of the test rubric	37
2.8.3 Characteristics of the input	38
2.8.4 Characteristics of the expected response	39
2.8.5 Relationship between input and response	40
2.9 Characteristics of test-takers	41
2.9.1 Personal characteristics	42
2.9.2 Topical knowledge	42
2.9.3 Affective schemata	42
2.9.4 Language skills	43
2.10 Speaking skill	45
2.11 Speaking and language acquisition theories	47
2.12 Testing speaking skills	50
2.12.1 Types of oral tests	52
2.11.1.1 Imitative tasks	52
2.11.1.2 Intensive tasks	53
2.11.1.3 Responsive tasks	54
2.11.1.4 Interactive tasks	55
2.11.1.5 Extensive tasks	56
2.12.2 Scoring procedures for measuring speaking proficiency	57
2.12.3 The assessment of speaking and language teaching methods	60
2.12.4 The importance of speaking skills and testing speaking	62
2.12.5 Major difficulties in testing speaking ability	64
2.12.6 Factors that affect speaking performance	65
2.13 Studies on testing speaking in EFL Context	67
3 METHODOLOGY	79

3.1 Research design	79
3.2 Research setting	79
3.3 Participants	81
3.4 Data collection	84
3.4.1 Mixed methods research design	85
3.5 Data collection tools	89
3.5.1 Questionnaire	90
3.5.2 Semi-structured interviews	92
3.5.3 Video-stimulated recall protocols	93
3.6 Data analysis	94
4 FINDINGS	97
4.1 Quantitative findings	97
4.1.1 Instructors' questionnaire results	97
4.1.2 Students' questionnaire results	99
4.2 Qualitative findings	101
4.2.1 Interviews with the instructors	102
4.2.1.1 Factors that affect students' speaking performance	102
4.2.1.1.1 Examinee-based factors	103
4.2.1.1.2 Rater-based factors	106
4.2.1.1.3 Context-based factors	107
4.2.1.2 Challenges that affect students' speaking	
performance	109
4.2.1.2.1 Examinee-based challenges	110
4.2.1.2.2 Rater-based challenges	112
4.2.1.2.3 Context-based challenges	113
4.2.1.3 Facilitators that affect students' speaking	
performance	114
4.2.1.3.1 Examinee-based facilitators	116
4.2.1.3.2 Rater-based facilitators	117
4.2.1.3.3 Context-based facilitators	118
4.2.2 Interviews with the students	120
4.2.2.1 Factors that affect students' speaking performance	120
4.2.2.1.1 Examinee-based factors	121

4.2.2.1.2 Rater-based factors	123
4.2.2.1.3 Context-based factors	124
4.2.2.2 Challenges that affect students' speaking performance from students' point of view	125
4.2.2.2.1 Examinee-based challenges	126
4.2.2.2.2 Rater-based challenges	127
4.2.2.2.3 Context-based challenges	127
4.2.2.3 Facilitators that affect students' speaking performance	129
4.2.2.3.1 Examinee-based facilitators	130
4.2.2.3.2 Rater-based facilitators	131
4.2.2.3.3 Context-based facilitators	132
4.3 Video-stimulated recall protocols	133
4.3.1 Video-stimulated recall interviews with the instructors	134
4.3.1.1 Factors that affected students' speaking performance	134
4.3.1.1.1 Examinee-based factors	135
4.3.1.1.2 Rater-based factors	136
4.3.1.1.3 Context-based factors	136
4.3.1.2 Suggestions for re-administering the exam	138
4.3.1.2.1 Context-based suggestions	140
4.3.1.2.2 Rater-based suggestions	141
4.3.2 Video-stimulated recall interviews with the students	142
4.3.2.1 Factors that affected students' speaking performance	143
4.3.2.1.1 Examinee-based factors	144
4.3.2.1.2 Rater-based factors	146
4.3.2.1.3 Context-based factors	146
4.3.2.2 What students would do if they were to retake the exam	149
5 CONCLUSION, DISCUSSION, AND SUGGESTIONS	153
5.1 Conclusion	153

5.2 Discussion	155
5.2.1 The role of affective factors	156
5.2.2 The role of readiness	159
5.2.3 The role of linguistic competencies	161
5.2.4 The role of content production	162
5.2.5 The role of exam raters	163
5.2.6 The role of contextual factors	164
5.3 Pedagogical implications	167
5.4 Suggestions	169
5.4.1 Suggestions for EFL educators	169
5.4.2 Suggestions for curriculum designers	170
5.4.3 Suggestions for test developers	171
5.4.4 Suggestions for future research	172

REFERENCES

APPENDICES

RESUME

LIST OF TABLES

	<u>Page</u>
Table 2.1 Task characteristics	35
Table 2.2 Distinguishing features of reactivity between input and response	41
Table 2.3 Micro- and macro-skills of oral production	51
Table 3.1 Number of participants during data collection	81
Table 3.2 Instructors' demographic information (Questionnaire)	82
Table 3.3 Students' demographic information (Questionnaire)	83
Table 3.4 Reliability results of piloting with teachers	91
Table 3.5 Reliability results of piloting with students	92
Table 3.6 Time course for data collection	94
Table 4.1 Instructors' overall questionnaire results	97
Table 4.2 Mean scores of the parts in the instructors' questionnaire	98
Table 4.3 Students Overall Questionnaire Results	99
Table 4.4 Mean scores of the parts in the students' questionnaire	101
Table 4.5 Factors that affect students' speaking performances from instructors' point of view	103
Table 4.6 Challenges that affect students' speaking performances from instructors' point of view	109

Table 4.7 Facilitators that affect students' speaking performances from instructors' point of view	115
Table 4.8 Factors that affect students' speaking performances from students' point of view	121
Table 4.9 Challenges that affect students' speaking performances from students' point of view	125
Table 4.10 Facilitators that affect students' speaking performances from students' point of view	129
Table 4.11 Factors that affected students' speaking performances from instructors' point of view	134
Table 4.12 Suggested changes for re-administering the exam	139
Table 4.13 Factors that affected students' speaking performances from students' point of view	143
Table 4.14 What students would do if they were to retake the exam	149

LIST OF FIGURES

	<u>Page</u>
Figure 1.1 Factors in performance assessment	2
Figure 1.2 Model of assessment of oral communication	3
Figure 2.1 Tests, assessment, and teaching	18
Figure 2.2 Usefulness	25
Figure 2.3 Criteria for testing a test	25
Figure 2.4 Practicality	26
Figure 2.5 Reliability	27
Figure 2.6 Authenticity	33
Figure 2.7 Some components of language use and language test performance	45
Figure 3.1 Mixed methods design matrix with mixed methods design shown in four cells	86
Figure 3.2 Basic mixed methods design	88
Figure 3.3 Research design: Explanatory sequential mixed methods	89
Figure 3.4 Creswell's Model of Coding Process	96

LIST OF ABBREVIATIONS

AFH: Affective Filter Hypothesis

AUSFL: Anadolu University School of Foreign Languages

COH: Comprehensible Output Hypothesis

EFL: English as a foreign language

ELT: English Language Teaching

IELTS: International English Language Testing System

IH: Interactional Hypothesis

L2: Second language

MM: Mixed-methods

MMR: Mixed Methods Research

SLA: Second Language Acquisition

SFL: School of Foreign Languages

ST: Sociocultural Theory

TOEFL: Test of English as a Foreign Language

TOEIC: The Test of English for International Communication

VSRP: Video stimulated recall protocol

WTC: Willingness to communicate

1 INTRODUCTION

For many years, speaking skills have been regarded as one of the most significant competences to be improved in foreign language learning. Bygate (1987) underlines the importance of speaking in both first and second languages. Certain approaches such as Communicative Language Teaching, The Natural Approach, Cooperative Language Learning, etc. have been developed in this direction (Richards & Rogers, 2002). In the 21st century, the importance of communicating while learning a language is still undeniable. However, oral skills are still quite challenging for many EFL and ESL learners (Alonso, 2018; Qasemi, 2020).

In addition to all these, the assessment of speaking skills has also become an important part of the language learning process. Many high-stake organizations such as Cambridge, Pearson and Educational Testing Service, make considerable use of speaking tests in their quite notorious tests, Test of English as a Foreign Language (TOEFL), International English Language Testing System (IELTS) and Pearson Test of English (PTE), to assess language skills. While some of these oral exams are utilized to determine the proficiency levels of English as a Foreign Language (EFL) learners, others are used at tertiary levels to measure mid-semester and end-of-semester achievement along with the other skills including reading, listening and writing. Also, in many universities where the medium of instruction is English, speaking has a significant proportion of the total language skills measured. However, when measuring this skill, it is considered that there are many factors that affect the student's achievement such as raters, pairs, criteria, test items, affective factors, etc.

1.1 Background to the Study

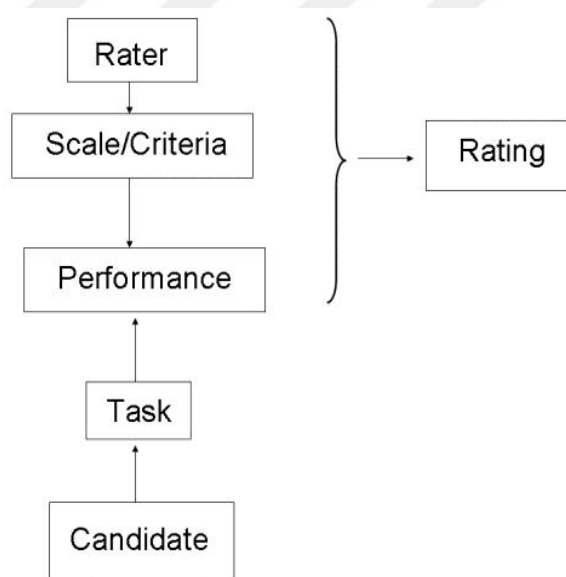
Speaking has been regarded as one of the most challenging skills to assess for the EFL teachers who are the raters of the testing process (Bachman & Palmer, 1996; Louma, 2009; Souri & Merç, 2019). The primary reason for this difficulty is that there are many variables involved in assessing speaking skills including raters' interpretations, the tasks or the individual differences of the examinees (Kitao & Kitao, 1996; Lado, 1961).

McNamara (1996) highlights three main issues that impact learners' language performance: the complexity of communicative competence, the role of contextual variables in assessment, and the limits of what can be tested. To address the first issue,

which is defining communicative competence, he proposes developing a comprehensive model that captures not only individual language ability but also the dynamic interaction between interlocutors. The second challenge concerns the influence of various contextual variables such as the type of task, the characteristics of participants, and the setting on assessment outcomes; thus, he emphasizes the need for empirical research on how important each variable including tasks, participants, settings, issues, scales in teachers' performance appraisal. The final issue involves setting realistic limits on what aspects of language use can be fairly and reliably tested. Namely, once a picture of the impact of these variables is formed, teachers need to decide on the boundaries for testability. In his much-quoted figure, McNamara (1996) highlights multiple elements that shape examinees' scores like rater assessing via a scale, which is different from traditional fixed-response assessment that involves only candidate-instrument interaction.

Figure 1.1

Factors in Performance Assessment



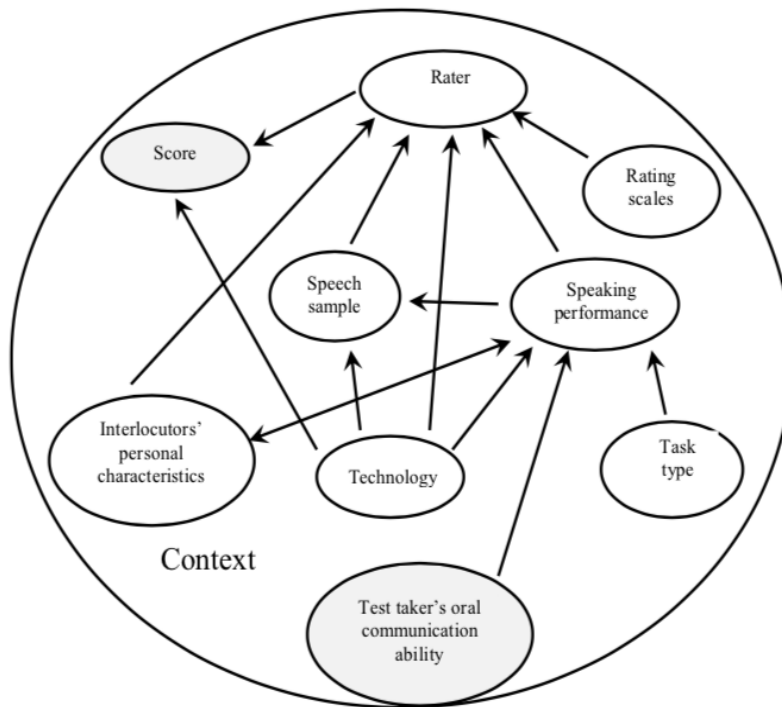
Reference: McNamara, 1996

According to Ockey and Li (2015), “assessing oral communication is a rather complicated process” since a series of factors make a contribution to test-takers’ scores (p. 2). As can be seen in Figure 1.2 below, these factors which might have an impact on scores during a test administration, include task type, other interlocutors’ personal characteristics,

technology used for the assessment, the actual speaking performance and resulting speech sample, rating scales, and raters.

Figure 1.2

Model of Assessment of Oral Communication



Reference: Ockey & Li, 2015

Namely, the review of relevant literature reveals that a number of factors are at play during the oral examination. Some of these can be further specified as follows.

Rater-based factors are one of the variables that require close attention during oral assessment of learners. Rater-based factors comprise rater cognition, rater characteristics and expertise. İnan (2019) states the most complex factor affecting the assessment of speaking is the rater cognition as the rating process happens in their minds and is uncontrollable. Previous research suggests that raters are subjected to different cognitive processes such as examining the tasks, monitoring the students, interpreting the criteria and making decisions before scoring (McNamara, 1996; Fulcher, 2003; Louma, 2009, etc.). It has also been cited in the literature that raters score differently for the same performance on the same task (Ahmadi & Sadeghi, 2016; Brown, Iwashita & McNamara, 2005; Chalhoub-Deville & Wigglesworth, 2005; Han, 2016). Thus, rater-related situations have been an issue in oral assessment.

Secondly, examinee-based factors also affect the scoring. These factors include affective factors, listening ability, being paired, and speaking problems. Affective factors are important components that have a notable effect on speaking assessment and influence language learning success or failure (Oxford, 1990). For example, Ur (1996) mentions inhibition, which affects students' state of anxiety (p. 121). Namely, some students may feel worried about making mistakes, fear being criticized, and become shy due to the attention that their speech draws. Yet, the affective factors not only include negative feelings such as anxiety, worry or stress, but they also incorporate motivation and self-efficacy, etc. For example, Krashen (1982) highlights the importance of motivation, self-confidence and anxiety in his Affective Filter Hypothesis, which places a very important role in second language acquisition.

In addition to affective factors, listening ability should be taken into consideration. Doff (1988) emphasizes the necessity of developing listening skills to be able to develop speaking skills as it is a prerequisite for students to understand what is said to them to have a successful conversation. Similarly, Shumin (1997) states that when someone is speaking, the other accompanies the conversation by means of the listening process. In a conversation, each speaker plays the role of a listener and a speaker, and one side is definitely incapable of responding unless they understand what is said. That is, speaking and listening are closely associated.

Furthermore, during most of the speaking tests, two examinees are paired and asked to implement a task together such as discussion, making arrangements, agreeing/disagreeing, picture description or gap-filling tasks. Thus, there might be an effect of pairing a weak and a strong student on the assessment. Salaberry and Burch (2021) highlight that paired speaking test formats are considered problematic for evaluating individual learners' second language proficiency as differences between interlocutors such as in proficiency level, familiarity, gender, and other variables might influence the results. Ur (1996), for example, states four basic problems with speaking activities, one of which is the "low or uneven participation" (p. 121). She underlines that this problem leads some learners to dominate while others produce very little or no language at all, which can also be observed in speaking exams.

In a similar vein, students' speaking problems such as active participation in the classroom along with the level of commitment out-of-the classroom and the knowledge

and the use of certain linguistic strategies may have an impact on the scoring. Liu (2022), in a preliminary article, developed a program named the Schema Association Training (SAT) to assist his learners to improve their discussion abilities in and outside the classroom as well as boost their confidence during speaking assessment. He believes SAT helps learners elaborate better in speaking assessments such as the Singapore-Cambridge General Certificate of Education (GCE) Ordinary Level (O-Level) English language oral communication exam and the University English Language Testing System (UELT) Speaking Test, etc.

Task-based cases can be regarded as the third variable, which include task types as well as the topic, topic knowledge, quality and difficulty might cause problems for both the raters and the examinees. As Fulcher (2000) states, performance-based test tasks should involve communication, and be authentic with a purpose and real-life-like communication. Another variable is the quality of the tests including validity, reliability and practicality. Task complexity, task difficulty and planning time can be regarded as the other variables that have an impact on the assessment process. Bachman and Palmer (1996) state that topical knowledge affects students' speaking performance because when a test-taker has already possessed the relevant topical knowledge, then certain test tasks will be easier for those compared to the ones who do not. Ur (1996) underlines four basic characteristics of a good speaking activity: *Learners talk a lot, Participation is even, Motivation is high and Language is of an acceptable level* (p. 120). The tasks in a spoken exam should include these characteristics too so that the quality and the difficulty of the tests can be arranged accordingly.

Criteria is another important variable found in literature in testing speaking. Speaking construct is a versatile conception that embodies several aspects like fluency, grammar, vocabulary, pronunciation, task achievement etc. (Brindley, 1991; Fulcher, 2003). Therefore, both the raters and the examinees should be informed about the rubric and the criteria beforehand so as to inform both parties about the learning outcomes (Fulcher, 2000). Besides, using a rubric also enhances the validity and the reliability of the speaking test (Jonsson & Svingby, 2007).

Ultimately, environmental conditions including the physical conditions, duration and the video-recording can also play a significant role during speaking exams. The physical quality of the exam room such as the lighting, the size, ventilation, etc. might affect the

raters and the examinees. Nation and Newton (2009) emphasize four types of performance conditions during an oral performance, one of which is time pressure. Furthermore, in some cases as in my institution, the speaking exams are video-recorded, which might affect both the scoring and the performance of the students during the assessment process.

In brief, testing speaking and scoring the examinees' performances involve many variables including raters, examinees, tasks, criteria and environmental factors that might have an impact on the scores of students' performances.

1.2 Statement of the Problem

It is important for language learners to improve their communicative skills, and therefore, assessing speaking skills gains importance, as well. However, assessing this skill is not a straightforward task for EFL teachers since it entails many details.

One of the main variables that affect students' speaking performance is the rater-based variables, mainly the rater cognition during the evaluation. For example, İnan (2019) examined the rater cognition of six non-native instructors at a state university. The participants were asked to perform and evaluate students' performances, which was followed by a verbal protocol analysis (VPA) to determine the harshness and leniency in scoring. The results showed both rubric-related, which included criteria in the rubrics, and non-rubric factors which were not found in the rubrics. That is, raters not only make use of the criteria in the rubrics but also use their own criteria in their minds about the tasks, the test-takers and themselves as raters.

Likewise, in his dissertation, Han (2020) investigated the effects of rater characteristics which are rater expertise, working memory capacity (WMC) and cognitive functionality with 90 raters from the US and the UK. The results yielded that rater expertise had a significant effect on raters' scoring accuracy though there were no significant interaction found between rater expertise and WMC. Besides, for the cognitive functionality, six of these participants reported their thinking processes in a cognitive lab session in order to disclose the mental mechanisms affecting raters' performances when scoring as well as the varieties in raters' strategy employment.

Examinee-based cases are the second factors that affect the scores of language learners. Qasemi (2020) examined the factors that affect students' speaking performance in an

English department with 108 language undergraduate learners. The results showed that many students perceive learning English as important and necessary, but have some speaking problems as they do not have the opportunity to practice outside the class, which influences their oral communication skills in English.

Besides, Jaya, Petrus and Pitaloka (2022) explored the factors affecting EFL university students' speaking performance and speaking problems using qualitative and quantitative tools. They categorized the speaking problems as affective-related including self-confidence and anxiety, socially-related like not having enough opportunity to practice this skill in the classroom, and linguistic problems such as lack of fluency, grammar, vocabulary and pronunciation. They also identified the factors that lead to speaking problems, some of which are lack of general knowledge, speaking practice, word usage and grammar practice as well as fear of mistake, low motivation, nervousness, shyness and fear of criticism.

Huang (2016) investigated the use of test-taking strategies with Test of English for International Communication Speaking Test (TOEIC-S) and their relationship with test performance of 215 Taiwanese EFL learners quantitatively. He reached three major conclusions. First, strategy use in the TOEIC-S test is diverse and includes more than one type of strategic behavior. Next, these strategic behaviors were matched to behaviors that enabled test takers to communicate both in real life and at work. Third, communication strategy use and cognitive strategy use both made a significant contribution to participants' TOEIC-S performance.

The third variable that might have an effect on students' oral performance is the task(s) in the tests. This may be related to not only the task type but also the testing mode, topic, topical knowledge, quality and difficulty, which might be troublesome for the raters and the examinees. In a study conducted by Tuan and Mai (2015) with Vietnamese students and teachers at a high school, the majority of the teachers stated that topical knowledge affects their students' speaking performance the most, which is followed by motivation and listening ability (which are regarded as examinee-based variations in this study). In a similar vein, 82% of the students stated that the topical knowledge is the highest influencing factor, which is followed by listening ability and confidence, respectively.

Alzahrani (2020) compared two oral proficiency testing modes: Oral Proficiency Interview (OPI) and Video Oral Communication Instrument (VOCI) regarding test-

takers' complexity, accuracy and fluency. OPI is a direct-testing mode whereas VOPI is indirect. The accuracy and the fluency level were found higher in OPI. It also explored the participants' perceptions on these two modes, and the results reflected a favor on OPI among participants since they felt more comfortable with this testing mode.

In addition, In'nami and Kouzami (2016) conducted a study to find out the percentage of variation in both L2 speaking and writing performance by mainly focusing on the tasks, raters and their interaction. For L2 speaking, they utilized 28 datasets from 21 studies, and the findings reflect that most of the score variances are defined by examinees' performance. Also, they deduced that task and task-related interaction effects explained more of the score variances than did the rater and rater-related interaction effects. However, they also concluded that context, scoring methods and criteria may vary performance on the tasks.

Criteria can be another factor that might have an impact on the scoring of oral performance. Karshi (2002) conducted research at Anadolu University School of Foreign Languages (AUSFL) with five EFL teachers on the use of two different analytical criteria while assessing 36 students' oral performances. The results of the study indicated a positive attitude towards four-band criteria rather than five-band one. The reason why they did not prefer the five-band criteria is that the descriptors were vague for them. Also, the interview results yielded some common problems about the four-band criteria, which are mainly about the descriptors not being clear enough, students' performance' incompatibility with the criteria, the problems of the bands, particularly with the highest and the lowest ones, and the big value differences among the bands. The researcher also suggests assessors to take training and attend norming sessions before implementing an oral test.

Another study, which highlights the importance of using clear criteria or rubrics in speaking assessment, was conducted by Babaii, Taghaddomi and Pashmforoosh (2016). The study examined the possible mismatches between students' and teachers' speaking assessment. In this regard, 29 students were asked to assess their audio-recorded performance in two different time intervals. They first rated themselves without receiving any assessment practice sessions or using criteria, and the second assessment was done after a practice session with criteria. At the same time, six EFL teachers rated these learners by using the same criteria. The results yielded a greater mismatch between the

scores of teachers and self-assessed learners without using criteria and practice which help reduce discrepancies between teacher and student evaluations by providing a shared understanding of performance expectations.

Khabbazzbashi and Galaczi (2020) delved into holistic, analytic, and part marking models (MMs) regarding their measurement properties and effect on candidates' Common European Framework of Reference (CEFR) classifications in an online speaking test conducted with 240 participants in two phases. In the first phase, the examinees' performance was evaluated holistically and by part. Depending on the findings of phase 1, in phase 2 the researchers compared the analytic and part MMs of 400 examinees. The findings indicated a significant effect of marking models. Specifically, in phase 1 around 30% of candidates were rated different CEFR level scores while in phase 2, it was found as 50%, which reflects the marking models impact on the participants speaking scores. Also, the researchers underpin the fact that higher scores were awarded to the candidates who were rated with holistic MM and lower with part MM.

Besides, examinees perform a speaking task under various conditions. Thus, environment-based conditions such as the physical conditions of the room, duration of the test and the condition of being-video recorded can play a role on the scores of the test-takers and the test-givers, as well. Tuan and Mai (2015) revealed performance conditions including time pressure is one of the factors which impacts students' speaking performance. That is, time limitation might lead the students to do less than optimally.

To conclude, when the previous studies were examined, there are a number of studies conducted to determine the factors that affect oral performance in various contexts, and many of them look into factors affecting speaking assessment from a limited perspective such as focusing on one factor. However, being as one of the most important skills of language learning and teaching, these factors deserve to be analyzed separately and deeply delving more into as many variables as possible. Specifically, in Türkiye, where there are thousands of university students learning EFL at School of Foreign Languages (SFLs), and where the speaking assessment owns a large proportion in the language achievement tests, it is highly important to examine the factors affecting the speaking exam results held in SFLs.

1.3 Significance of the Study

In the context of growing emphasis on communicative competence in foreign language education, this study addresses a key area: the variables that shape learners' speaking performance in oral achievement tests. Among four skills, speaking is probably the most complex and anxiety-inducing skills for language learners, particularly in test settings where pressure is high. By exploring the specific factors that influence learners' speaking performance during achievement tests, this study contributes to a deeper understanding of how emotional, linguistic, and contextual variables interact in speaking exam settings.

When the relevant literature is examined, it can be said that it lacks a thorough analysis of the factors that have an influence on the examinee's oral performance. Previous studies have predominantly focused on isolated factors, often from a single stakeholder group, such as test-takers or assessors (e.g., Alzahrani, 2020; Han, 2020; İnan, 2019; Jaya, Petrus & Pitaloka, 2022; Qasemi, 2020). Yet, the other possible factors that affect students' speaking exam performances might need to be delved in-depth by careful examination of the views of all three parties that make up the evaluation, which are the instructors, the students and the test-designers. Likewise, in the Turkish context, these factors were not fully investigated although many School of Foreign Languages (SFLs) in Türkiye perform speaking exams as an important part of their assessment. Therefore, it can be said that, there is a notable gap in the literature regarding a more holistic understanding of what shapes speaking performance in oral achievement tests. This study seeks to fill that gap by exploring the interplay of emotional, linguistic, rater-related, task-related, and contextual variables from multiple viewpoints.

At a practical level, the findings of this study can inform the design and administration of speaking assessments in Schools of Foreign Languages (SFLs) in Türkiye and in similar EFL contexts where standardized testing environments gain importance. Insights into learners' affective responses, task types, examiner behavior, peer effects, and environmental conditions can guide both educators and test developers in designing more valid, equitable, and supportive testing environments. Moreover, by raising awareness of these influencing elements, the study contributes to improved assessment literacy among instructors and test designers, and promotes more effective preparation strategies for learners. Ultimately, the research supports a more comprehensive and fair approach to

speaking assessment, addressing not only linguistic ability but also the psychological and situational factors that shape performance.

Finally, this research highlights the need for a more holistic approach to speaking assessment, considering not just language proficiency but also the psychological and situational dimensions of examinee performance. As such, the study has the potential to contribute to improved assessment literacy among educators and to enhance the overall fairness and reliability of speaking tests in EFL settings.

1.4 Purpose of the Study

The purpose of this study is to investigate the factors that influence EFL learners' speaking performance in oral achievement tests, as perceived by language instructors including classroom instructors, test designers, and administrators, along with EFL students. While existing research has examined some of these variables individually, a comprehensive analysis that brings together the perspectives of all stakeholders remains lacking, particularly within the Turkish EFL context.

This study aims to address that need by identifying and analyzing the variables that contribute to learners' speaking exam scores in a tertiary-level EFL setting. By doing so, the research intends to generate a deeper understanding of the components that impact oral language assessment and to support more effective teaching, testing, and learning practices.

Therefore, to my humble mind, to date, what factors affect students' speaking performances through the instructors', who are also assessors and test designers, as well as examinees' point of view has not been probed hitherto in depth. To this end, it is important to find out the factors that have an impact on EFL learners' speaking performance in achievement tests in order to provide awareness for tertiary level EFL learners to improve their oral competence and, ultimately, to take this competence further. Similarly, the instructors, test-designers and the examiners will be aware of the factors that impact the scores so that they can prepare and administer the speaking exams accordingly by paying close attention to these factors.

The purpose of this current study is to find out the factors that affect the EFL learners' speaking exam scores in their achievement tests from the teachers', students' and test-designers' perspectives. Bearing these in mind, the following research question arises:

1. What are the factors that affect EFL learners' speaking performance in the achievement tests from the point of view of:

- instructors?
- students?

1.5 Definitions of Key Terms

English as a foreign language (EFL): EFL refers to the teaching and learning of English in a country where English is not the primary language spoken. It typically involves learners studying English in their own non-English-speaking countries, often as part of their school curriculum or through private language courses, with limited exposure to real-life English use.

Language assessment: Language assessment is the process of evaluating a learner's language proficiency, skills, and progress. It involves the use of various tools and methods such as tests, quizzes, observations, and performance tasks to measure abilities in areas including listening, speaking, reading, and writing.

Language testing: Language testing is the process of designing, administering, and interpreting tests that measure a person's ability to use a language. It is important in language education by assessing learners' proficiency in areas such as speaking, listening, reading, writing, grammar, and vocabulary.

Examiner: An examiner is a professional who is responsible for administering, scoring, and evaluating language tests or assessments. They ensure that the testing procedures are followed correctly and that the results accurately reflect the test-takers' abilities.

The terms *tester*, *rater*, and *examiner* are used interchangeably throughout this dissertation to refer to individuals involved in test administration and evaluation.

Skill: A skill is the ability to perform a task well, usually gained through learning, practice, and experience. It involves applying knowledge and techniques to accomplish specific activities effectively and efficiently.

The terms *skill*, *ability* and *competency* will be used interchangeably to represent the capabilities required to perform specific tasks or functions.

Oral production: Oral production refers to the ability to express thoughts, ideas, and information verbally in a second or foreign language. It involves speaking skills,

including fluency, accuracy, pronunciation, vocabulary use, grammatical control, and discourse organization.

The terms oral production and oral performance are used interchangeably when they refer to how a student actually spoke during a test or presentation.



2 LITERATURE REVIEW

It is undeniable that testing has an important place in the language learning process. There are various reasons for language testing, including providing feedback on and self-evaluation of learners' both general and specific language competencies and abilities. Depending on the purpose for testing, the test types and the testing approaches vary. For example, a proficiency test is intended to evaluate test-takers' general language competencies with no obligation to follow a curriculum or course objectives, whereas an achievement test is in direct relation with the course content and objectives since it aims to not only assess the test-takers' success but also evaluate the course itself.

On the one hand, some scholars are in favor of integrative test methods, which rely on the unitary trait hypothesis, and assert that language proficiency is "indivisible". They view the four language skills as well as language systems (vocabulary, grammar, phonology) cannot be separated from each other. On the other hand, others find the unitary trait position very wrong because learners' background such as native country, field of study, educational status may have an impact on learner's performance in different language skills.

Besides, when testing each skill separately, speaking is generally found to be the most difficult one to assess (Brown, 2004; Louma, 2004). This challenge may stem from its complex and dynamic nature. The first difficulty may be related to its being simultaneous and dependent on the addressee's oral contributions. That is, oral communication necessitates real-time production as well as comprehension, which has limited processing time for the speaker. Plus, speaking "involves the simultaneous deployment of linguistic systems at several levels: phonemes, morphemes, the lexicon, syntax, and discourse (Bailey, 2020, p. 79). Test-takers are expected to produce accurate, appropriate, and intelligible spoken responses, which creates pressure on learners due to the time constraints during speaking tasks (Fulcher, 2014). In addition, Fulcher (2014) also emphasizes the demand on testers' valid evaluation of learners' true ability. A recent study conducted by Tavakoli and Skehan (2020) has also displayed the cognitive load in real-time speech production, which made the assessment procedure even more complicated.

Along with simultaneousness, speaking involves both linguistic (vocabulary, grammar and pronunciation) and paralinguistic components including stress, rhythm and

intonation, which makes the process more complicated. Louma (2004), for example, mentions the difficulty in assessing one's real oral proficiency since developing valid and reliable assessment criteria that are appropriately aligned with the targeted language skills is not simple. In a similar vein, according to Isaacs and Trofimovich (2019), developing more comprehensive assessment tools which can grasp the nuances in oral production is needful. Beside assessment tools, Brown and Abeywickrama (2019) further argue for the effects of linguistic complexity in the design of speech assessments.

Examiner-based factors can be another issue in assessing speaking. Since speaking is an interactive skill that requires interlocutors interacting in a specific context, this interaction between the speakers can have an impact on the flow of the conversation and the language type used. Bachman and Palmer (2010) assert that there is variety in mutual conversation, possibly influencing the test-taker's behaviors and responses, which, in return, can cause lack of standardization and inconsistencies among raters and scoring during assessments. Plus, the subjectivity during the evaluation of a speech poses another issue. Oral skill assessments are often regarded as subjective because it depends on the evaluators' judgments. Even though using a rubric or a rating scale is common among assessors, raters' biases and perceptions can have an impact on the application and the interpretation of these assessment tools. That's why Weir (2005) highlights the importance of inter-rater reliability to prevent inconsistencies. Rater training programs as well as automated scoring systems may be solutions to reduce subjectivity and increase reliability of these evaluations (Knoch & Elder, 2013).

Examinee-based factors including test anxiety and performance pressure can be deemed as other significant factors that cause difficulty in oral assessment. To exemplify, test anxiety, which can emerge from the fear of making mistakes, being shy, being judged by others, etc. may negatively affect the tester's performance. Horwitz, Horwitz, and Cope (1986) assert that anxiety can be a reason for learners to perform less well than their actual performance, and can lead to inaccurate assessment. Relatively recent research by Dewaele and MacIntyre (2019) examine the relationship between language anxiety and positive psychology, which focuses on human strengths, well-being, and optimal functioning (Seligman & Csikszentmihalyi, 2000). The results yielded that positive psychology reduces language anxiety and enhance speaking performance.

In the following part, the literature is compiled around tests, testing, speaking and testing speaking.

2.1 What is a test?

“Testing is an important part of every teaching and learning experience” (Madsen, 1983, p. 3). Given its significance, it is essential to define what a test is. A test is “a method of measuring a person’s ability, knowledge, or performance in a given domain” (Brown, 2003, p. 3). For Bachman (1990), a test is a type of measurement which tries to yield a person’s “specific sample of behavior” (p. 21). Hedge (2000) defines testing as “specific procedures that teachers and examiners employ to try to measure ability in the language” (p. 378). Shohamy (2001) perceives testing as a professional field which involves strict rules and applications. That is to say, testing has certain characteristics. First of all, a test is a method that includes techniques, procedures or items on which the test-taker has to execute (Brown, 2003). Shohamy (2001) emphasizes the objective type items in traditional testing, for these types of items lessen statistical unreliability. The second component of a test is that it should measure either the general ability of the language learners as in proficiency tests or certain competencies or course objectives as in achievement tests like quizzes or midterm exams (Brown, 2003). Third, it is vital for the test designers to be familiar with the test-takers since “tests measure individuals’ ability, knowledge or performance” (Brown, 2003, p. 3). In other words, although a test is the evaluation of the learner’s performance, it indicates the test-taker’s language ability, or competence. Lastly, a test focuses on a domain, which can be the overall language proficiency or more particular requirements as in a vocabulary test. In brief, Brown (2003) emphasizes that a test is an instrument to measure the test-taker’s ability within a specific domain.

2.2 Reasons for Testing

Tests are designed and administered for different purposes in education (Kaçar, 2023). Depending on the purpose of the test, there are many reasons why language teachers test the learners. Heaton (1990) underlines seven different reasons for testing. First, and maybe the most important reason is to find out how well the test-takers have mastered the language skills that have been recently taught. Thanks to these tests, learners have the opportunity to look back at what they have(not) done and evaluate their progress. The second reason is that tests encourage learners by showing their progress during language

acquisition. Particularly, the classroom tests can help them see their achievement in course goals and objectives. Another important function of tests is that they can help learners diagnose their learning difficulties and become aware of their weaknesses so that they can overcome these problems. Other reasons are associated with different purposes of test types. These include the learners' overall achievement, placing them in suitable levels, selecting them, and finding out about their proficiency level.

In a similar vein, Brown (2003) draws attention to the importance of tests as the learners can achieve opportunities to engage with authentic language use, be motivated by setting clear goals, and receive feedback which guides their learning progress. In addition, tests are one of the significant components of the learning environment as well as a strong curriculum.

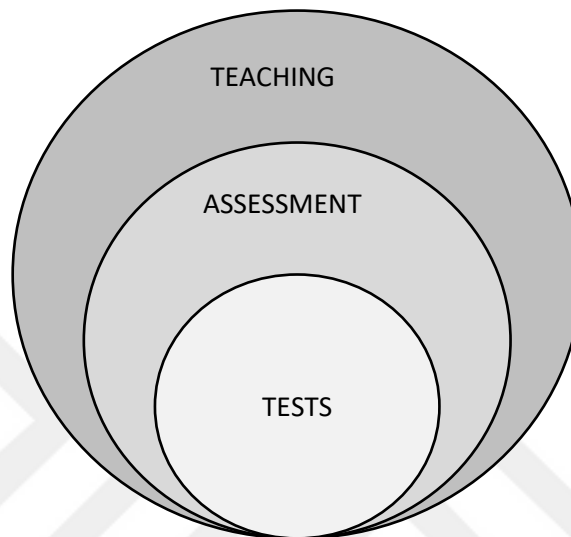
2.3 Testing and Assessment

Although the term “testing” has been viewed as old-fashioned and was preferred in the past (Ölmezer-Öztürk, 2018), and has been shifted to the term “assessment” (Inbar-Lourie, 2008), it is still widely used in the teaching-learning environments. Despite the fact that many scholars think they are synonymous, some scholars think the opposite. For Bachman (1990), a test is a type of measurement which tries to yield a person's “specific sample of behavior” (p. 21). Brown (2003) stated that the two terms are actually different. According to him, tests are designed for administrative reasons, and learners try to do their best to show their utmost performance “knowing that their responses are being measured and evaluated” (p. 4). Hedge (2001), for instance, states that assessment is a broader concept and a part of teaching and learning process in which learners progress can be in language learning can be monitored over time. She also adds that assessment is a “multi-faceted” concept while testing is “bolt-on procedure at end points in a learning program” (p. 376). Namely, tests are administered by teachers and examiners and intend to measure the ability in the language. On the other hand, assessment is defined as a longer continuum with more domains. Purpura (2016) defines assessment as “a procedure for eliciting test and non-test for the purpose of making inferences” (p. 191). Namely, assessment is a broader term which encompasses testing; thus, tests are regarded as “a subset of assessment” (Brown, 2003, p. 4). While tests are prepared tasks, assessment is a more general concept. To exemplify, a learner's responses to a question or his/her

comments in the classroom are considered as assessment. Below is the diagram of the relationship among teaching, testing, and assessment.

Figure 2.1

Tests, Assessment, and Teaching



Reference: Brown, 2003, p. 5.

2.4 Types of Assessment

In order to better understand the differences between tests and assessment, the types of assessment should be revisited. Based upon the purpose of assessment, the types of assessment that takes place in the literature are as:

- informal and formal assessment (Brown, 2003)
- formative and summative assessment (Brown, 2003; Hedge, 2001)

Brown (2003) states that informal assessment, which can be incidental and unplanned, can take various forms such as using appraisals after a classroom task or putting a smiling emoji on a written work. *Informal assessment* is commonly seen in classroom tasks where students display their performance without any recorded results or rigid decisions on students' language competence. On the other hand, in *formal assessments*, the major aim is to identify the learners' skills and knowledge via specifically designed exercises or tasks. Therefore, they are planned and systematic, and provide feedback and appraisal to both the teacher and the learner (Brown, 2003). Brown (2003) also emphasizes that “all

tests are formal assessments, but not all formal assessment is testing” (p. 6) as in the example of student portfolios which is regarded as formal assessment but not as a test.

Brown (2003) and Hedge (2001) distinguished assessment into two categories depending on their function: *formative* and *summative*. In formative assessment, the primary intention is to help learners continue their language development through pertinent feedback. Formative assessment is a tool for language teachers to design further classroom work depending on the information obtained from these assessments, and diagnose students’ strengths and weaknesses individually (Hedge, 2001). According to Harlen and James (1997), formative assessment is regarded as a part of the instructional process, which enables both instructors and learners to understand when to do adjustments in the teaching-learning process nature (Garrison & Ehringhaus, 2007). For Brown (2003), all informal assessments are formative. In summative assessment, on the other hand, the purpose is to assess what has been learned at the end of a course or instruction. In other words, students have the opportunity to look back and evaluate how well they have accomplished the course objectives. Besides, the assessment tasks are not necessarily prepared by the class teachers. On the contrary, summative assessment is usually imposed by an institution or a ministry (Hedge, 2001). That’s why summative assessments are generally high-stake and formal exams in nature (Garrison & Ehringhaus, 2007). Some typical examples for summative assessment are quizzes, midterm and final exams in a course or semester.

Besides these typical assessment types including formal-informal and formative-summative dichotomy, modern education also includes alternative assessments, which evaluate students via methods other than traditional standardized tests (Birenbaum & Dochy, 1996). The term alternative assessment can be cited under various other terms such as authentic, holistic, non-traditional or integrative assessments. Wiggins (1990), for example, uses the term authentic assessment to describe real-life tasks that require learners to use their knowledge in meaningful contexts. The primary goal of this type of assessment is to enhance student learning (Gulikers, Bastiaens & Kirschner, 2004) through construction of meaning (Mueller, 2005); thus, typical alternative assessments are generally associated with constructivist and learner-centered approaches. The emphasis is on the process rather than final outcomes. Therefore, the alternative

assessments include performance-based assessment and peer/self-assessments including portfolios, presentations, and projects.

2.5 Kinds of Language Tests

Since the tests are implemented to obtain information about test-takers' competency, different kinds of tests have been developed depending on the information being sought. The different kinds of language tests are:

- proficiency tests
- achievement tests
- diagnostic tests
- placement tests (Hughes, 2020)

The main purpose of *proficiency tests* is to determine the general language ability of the test-takers so as to be deemed as proficient regardless of getting training (Hughes, 2020). For Hughes (2020), proficiency tests vary in content and level of difficulty; however, they share one common characteristic: they are not based on the language courses that test-takers have previously finished. Cambridge examinations such as the First Certificate Examination and the Proficiency Examination are typical examples of proficiency tests. Harmer (2001) exemplifies that individuals take proficiency tests when they want to get into a foreign university, get a job, or certificate.

Contrary to the proficiency tests, *achievement tests* assess test-takers' success individually and as a group or evaluate whether the courses have met their learning objectives, so are directly associated with the course content (Hughes, 2020). Harmer (2001) also calls it progress test since their aim is "to measure learners' language and skill progress in relation to the syllabus they have been following" (p. 321). Achievement tests are limited to assessing particular learning goals or objectives (Brown & Abeywickrama, 2010). Regarding this, Hughes (2020) indicated two types of achievement tests, which are final achievement tests and progress achievement tests. Final achievement tests, as their name suggests, are administered at the end of a course, and can be developed by the ministries of education or the members teaching at the institution. Final achievement tests are also different from proficiency tests since the former ones should be dependent on the objectives of a particular course while the latter focuses on the general proficiency level. Like final achievement tests, progress achievement tests also rely on the course

objectives, yet “are intended to measure the progress that the students are making” (Hughes, 2020, p. 12). However, in contrast to final achievement tests, progress achievement tests are based on short-term objectives. Pop quizzes and midterm exams are the typical examples to progress achievement tests (Hughes, 2020).

Diagnostic tests are intended to diagnose a particular aspect of language (Brown, 2007), and to find out the learners’ strengths and weaknesses (Bachman & Palmer, 1996; Hughes, 2003). According to Harmer (2001), diagnostic tests help teachers discover learner difficulties, deficiencies, and gaps in their knowledge so that teachers can “do something about them” (p.321). These tests are generally implemented before a course to identify what further teaching will be. Thanks to these tests, students are able to detect the gaps in the target language acquisition, and overcome their weakness through practice and examples of the sources of information. A diagnostic test of English grammar that aims to assess learners’ understanding of key grammatical structures such as verb tenses, sentence formation, and agreement rules can be a good example of a diagnostic test. By pinpointing specific areas where students struggle, such a test helps instructors tailor instruction to meet individual learning needs effectively.

Placement tests are utilized to gain information about the learners’ abilities in order to place them in the most appropriate teaching program or level (Hughes, 2020). Placement tests frequently assess students’ grammar and vocabulary knowledge along with productive and receptive skills, which will be followed and used during the education program (Harmer, 2001). For example, language schools or schools of foreign languages carry out placement tests at the beginning of the academic year to assign the learners to appropriate classes at various levels.

At Anadolu University School of Foreign Languages (AUSFL), three primary types of assessment are administered: proficiency, achievement and placement tests. Proficiency and placement tests are conducted at the commencement of each academic year. The proficiency tests are developed by the Proficiency Exam Unit members in alignment with the designated course outcomes. Students who demonstrate proficiency are exempted from language courses at SFL, and proceed directly to their departmental studies, whereas those who do not meet the proficiency criteria undertake the placement test to be assigned to the appropriate language level. Achievement tests, on the other hand, are prepared by the Mid-Term Exam Unit, which holds primary responsibility for preparing exams for

each level in coordination with the Program Development and Material Support Units, ensuring that the assessments correspond with the syllabus and the learning outcomes organized by the Program Development Unit. To ensure that students are familiar with the question types, can practice beforehand, and are not faced with any surprises during the exam, sample questions are available on the AUSFL website (<https://ydyo.anadolu.edu.tr/ders-materyalleri/sinav-ornekleri>, and <https://ydyo.anadolu.edu.tr/ders-materyalleri/yeterlik-sinavi>)

2.6 Kinds of Testing

This part is an attempt to explain the two possibly contradictory approaches to test construction, as set forth below:

- norm-referenced versus criterion-referenced testing
- discrete-point versus integrative testing
- direct testing versus indirect testing
- objective testing versus subjective testing (Brown, 2004)

Norm-referenced tests calculate the mean, median, standard deviation and/or percentile rank of each test-taker's score (Brown, 2004). The aim is to rank the test-takers in a mathematical order, and their scores are announced numerically (such as 80 out of 100), and in percentages (such as 80 percent). Norm-referenced tests are standardized tests like the Test of English as a Foreign Language (TOEFL), and need to have fixed answers to be able to score immediately. *Criterion-referenced tests*, however, provide feedback to the test-takers usually via grades (Brown, 2003). Classroom tests are good examples of criterion-referenced tests. Therefore, the teacher workload can be heavier compared to the norm-referenced tests since the teacher supplies feedback to students.

Discrete-point tests are formulated by disintegrating the language components into parts, and testing these parts separately. These components can be the language skills which are listening, speaking, reading and writing as well as the subskills that are grammar, vocabulary and pronunciation. That is, discrete-point tests test only one thing at a time such as choosing the correct verb tense (Harmer, 2001). Brown (2003) urges upon the fact that this type of test is generally in decontextualized form which can cause confusion among test-takers. Therefore, since the emphasis was put on the communication, context

and authenticity, the need for new testing approaches have emerged. *Integrative testing* has thereby arisen. Students need to use a variety of language items at a time in integrative tests (Harmer, 2001). Cloze tests and dictations are the quintessential examples to integrative tests. In cloze tests, the test-taker is required to read a text where certain words, generally the sixth or seventh words are omitted, and fill in the gaps with an accurate word, which demands the test-taker to have competence in not only knowledge of vocabulary, grammar and discourse but also reading skills and strategies so that the test-taker is able to predict the word that fits into the next sequence (Brown, 2003). In a similar vein, dictation requires more than one language competency including listening ability, writing ability, the use of correct spelling and grammatical and discourse competencies.

Direct testing, as the name implies, focuses on measuring the test-taker's performance on a skill that is meant to be measured. The testing of productive skills is an example of direct testing where, for example, the testers utilize a speaking test to measure how well a test-taker pronounces the target language. According to Harmer (2001), when a test item asks examinees to perform the communicative skills, then it is direct. Direct testing is easier to perform for the productive skills tests than the receptive ones as in listening and reading tests the test-taker also requires to answer the test items accurately (Hughes, 1989). In addition, direct testing is straightforward in terms of application and assessment of the competency in a skill. Besides, a positive washback effect can be created by this type of testing because "practice of the test involves practice of the skill" (Hughes, 1989, p. 15). *Indirect testing*, on the contrary, does not attempt to test skill directly but seeks to measure the latent abilities. Harmer (2001) identifies indirect testing as a test which measure examinees' knowledge and ability by trying to expose what is beneath their productive and receptive skills. To illustrate, in one part of the TOEFL exam, test-takers identify the incorrect word/chunk in a sentence where four words/chunks are underlined. The purpose is to test the taker's writing ability in a roundabout way. Nonetheless, whether indirect testing fully measures the skill which is meant to be tested via underlying abilities is debatable.

Objective testing approach is merely about test scoring. That is, when no judgment is called for by the scorer, it is said to be objective as in the multiple-choice tests, matching and true/false questions (Hughes, 1989). In an objective test, each item has only one accurate answer, which makes it more reliable and easier to assess. On the contrary, when

the scorer makes any judgments on the test performance, then it is called *subjective*. Compositions, e-mails, discussions and talks are examples of subjective testing. Heaton (1990) deems subjective tests are superior to the objective ones as they “offer better ways of testing language skills and certain areas of language” (Heaton, 1990, p. 32). Nevertheless, although subjective tests give much freedom and flexibility to the test-takers, the scorers need to be competent enough to make fair judgements and mark the tests reliably.

In sum, depending on the test objectives, tests can appear in various forms including norm-referenced vs criterion-referenced, discrete-point vs integrative, direct vs indirect, objective vs subjective, traditional vs alternative testing besides communicative language testing, performance-based assessment and computer-based testing.

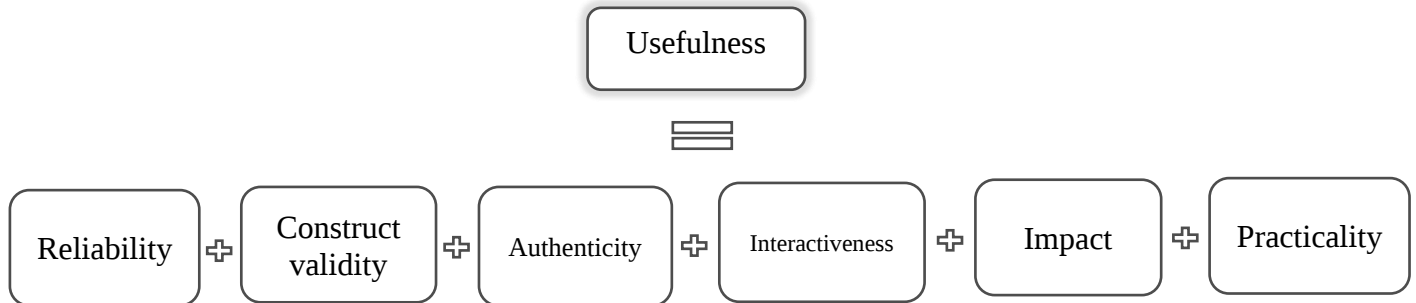
At AUSFL, a diverse range of testing types is employed in both achievement and proficiency tests to assess students' language proficiency. The school applies criterion-referenced testing, for example in achievement tests such as quizzes and midterm exams, to measure learners' proficiency against predefined standards and provide feedback to the test-takers via grades. Both discrete-point, focusing on isolated language elements such as grammar or vocabulary, and integrative tests, which assess learners' ability to combine these elements in meaningful communication, are utilized in achievement and proficiency tests. Moreover, direct testing methods are incorporated to evaluate actual language use through tasks like speaking and writing, whereas indirect tests measure underlying language skills through receptive tasks such as multiple-choice grammar or vocabulary items. AUSFL also balances objective testing, which includes clear right-or-wrong answers like multiple-choice or true/false items, with subjective testing, where responses are evaluated based on quality, such as essays and oral presentations. Sample questions types are available on the AUSFL website (<https://ydyo.anadolu.edu.tr/ders-materyalleri/sinav-ornekleri>, and <https://ydyo.anadolu.edu.tr/ders-materyalleri/yeterlik-sinavi>).

2.7 Principles of Language Testing

Bachman and Palmer (1996) investigate the qualities of language tests under the title of test usefulness. To them, in order for a test to be useful, several qualities contribute to the overall usefulness of a test, “all of which contribute in unique but interrelated ways”, as displayed in Figure 4 below.

Figure 2.2

Usefulness

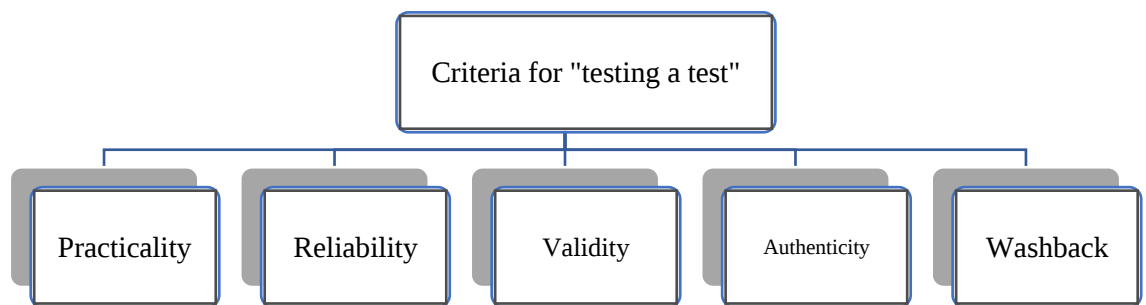


Reference: Bachman & Palmer, 2002

Brown (2003) has addressed this issue from a similar perspective, and clustered these principles for “testing a test” under five headings: practicality, reliability, validity, authenticity and washback, which are specified below (p. 19).

Figure 2.3

Criteria for testing a test



Reference: Brown, 2005

2.7.1 Practicality

According to Brown (2003), “An effective test is practical.” (p. 19). Weir (1993) states it as “The test must be practicable” (p. 52). Practicality includes “the context for the implementation of the texts” (Weir, 1993, p. 51). Namely, test usability should be taken into consideration. A test should be fairly low-cost, remain within suitable time limits, not be difficult to carry out, and have specific and time-saving scoring policy (Brown, 2003). Likewise, Bachman and Palmer (1996) describe practicality as the association

between the resources necessary for the test design, development and use, and the available resources. That is, practicality concerns whether the test specifications can be met with the existing resources. They represent practicality in the figure below (Bachman and Palmer, 1996).

Figure 2.4

Practicality

If practicality ≥ 1 , the test development and use is practical.

If practicality < 1 , the test development and use is not practical.

$$\text{Practicality} = \frac{\text{Available resources}}{\text{Required resources}}$$

Reference: Bachman & Palmer, 1996, p. 36.

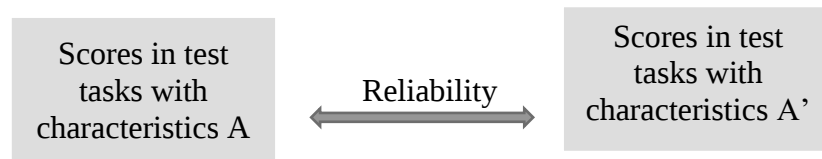
Harris (1969) also emphasizes the practical considerations of a test. First of all, a test should be economical both in terms of money and time. That is, a test should not be too expensive, but at the same time, time for administering and scoring the test should be taken into consideration. Besides, he highlights the ease of administration and scoring since the administration process and the scoring procedure can have significant effects on the practicality of the test. Finally, ease of interpretation is another concern that deserves attention. Harris (1969) states that the testers should have general guidance and implement the test accordingly.

2.7.2 Reliability

Harris (1969) defines it as “the stability of test scores” (p. 14). Brown (2003) puts it a bit differently, and asserts that “A reliable test is consistent and dependable” (p. 20). According to Harmer (2001), “a good test should give consistent results” (p. 322), making it reliable. Bachman and Palmer (1996) describe reliability as “the consistency of measurement” (p. 19), as in Figure 2.5 below.

Figure 2.5

Reliability



Reference: Bachman & Palmer, 1996, p. 20.

Based on the aforementioned definitions, reliability can be summarized as: if a test is carried out again with the same group of students, it is expected to obtain similar results. Or, as represented in Figure 2.5, when a test-taker takes either test A or test A' interchangeably, s/he should obtain a similar score no matter which test s/he takes. However, lack of reliability, or in other words – unreliability, can be due to more than one factor. These factors are student-related reliability, rater-reliability, test administration reliability and test reliability (Brown, 2003).

Student-related reliability issues are generally due to physical reasons such as illness, fatigue or psychological reasons like anxiety, which eventually causes deviation of the test-taker's "true" score from the "observed" one (Brown, 2003 p. 21). Mousavi (2002, p. 804, as cited in Brown, 2003) added "test-wiseness", the use of effective test taking strategies, as another student-related reliability factor.

Weir (1993) and Harmer (2001) highlight the importance of rater reliability due to its overall effect on test reliability. Thus, some rater-related factors can come into play such as human error and subjectivity. Brown (2003) categorizes these rater-related reliability factors into two groups: inter-rater reliability and intra-rater reliability. Inter-rater reliability happens when two scorers' give inconsistent marks for the same test. This situation can be a result of inexperience, misuse of criteria, and biases. Intra-rater reliability occurs when a scorer, for example, has to score a big number of tests, which can lead to inconsistency among the evaluation of the students' tests such as being harsher to the first students' papers than the later ones. This can be prevented by reading the first half of the papers before yielding any ultimate scores, and then reading through the whole papers (Brown, 2003).

The test administration conditions are also important (Brown, 2003). For instance, the noise coming outside the classroom may result in comprehension difficulties. Or the

physical conditions of the classroom such as inadequate lighting, ventilation, temperature, etc. can lead to reliability issues.

Unreliability may also be a result of the test itself (Brown, 2003). To exemplify, when a test is too long or when there is time pressure, the test-taker's performance may be affected adversely. Besides, poorly written test items can cause unreliability because they may confuse or mislead test-takers, leading to responses that do not accurately reflect their true knowledge or skills. Ambiguous wording, unclear instructions, or overly complex phrasing can cause misunderstandings, resulting in inconsistent answers that vary not due to ability but because of the item's flaws.

In brief, reliability is an essential quality, and therefore, the test scores should be consistent. However, it may not always be possible to discard all inconsistencies in a test, but it is wise to try to lessen the impacts of these potential factors, thereby amplifying the overall reliability of the test results.

2.7.3 Validity

In addition to reliability, validity is also of great importance in test design. Bachman (1990) remarks that the fundamental consideration in a test development and use process should be to dwell on both the reliability of the test scores and also the validity of the interpretations and uses of test scores. Namely, it has become prominent to ensure not only that a test produces consistent results but also that the test truly measures what it is intended to measure and that the interpretations and decisions based on the test scores are appropriate and meaningful.

Many definitions have been made to explain validity. For Henning (1987), validity is the “appropriateness of a given test or any of its component parts as a measure of what it is purposed to measure” (p. 170). Hughes (1989) stated that “a test is said to be valid if it measures accurately what it is intended to measure” (p. 22). Harris (1969) drew attention to two questions that are necessary to be searched answered for validity: (1) what precisely does the test measure?, and (2) how well does the test measure? (p. 19). Gronlund (1998) described it as “the extent to which inferences made from assessment results are appropriate, meaningful, and useful in terms of the purpose of the assessment” (p. 226). Weir (1993) simply puts validity as “A test should test what the writer wants it to test. It should be valid.” (p. 46). Harmer (2001) states that “a test is valid if it tests what it

is supposed to test” (p. 322). Although many definitions of validity exist in literature, in its broadest sense, it refers to whether a test actually measures what it is intended to measure. Thus, for Fulcher and Davidson (2007), for example, validity is “the central concept in testing and assessment” (p. 3). Similarly, Bachman (1990) sees validity as the most important quality of test interpretation and use. Brown (2003) thinks it is the most complex and the most important testing principle. To give a simple example, if test aims to test the listening ability, then it has to measure the listening ability. On the other hand, if a tester asks the test-takers to write the words they hear in a listening text, though it may seem practical to administer and reliable in terms of scoring, it may not constitute a valid test since it lacks comprehension.

Validity is broken down into five types, which are:

- content-related validity,
- criterion-related validity,
- construct-related validity,
- consequential validity and
- face validity (Hughes, 1989)

One of the most commonly cited type of validity is *content validity*. For Hughes (1989), if a test covers “proper sample of relevant structure”, then it has content validity (p. 22). Harris (1969) specifies that when a test is designed to measure a particular skill or reflect the content of a specific course, it should be based directly on that skill or course outline. Furthermore, the test items should be decently distributed, and care should be taken to avoid focusing disproportionately on items that are easiest to construct or most familiar to the test developer, as this can distort the overall representation of the content. Harmer’s (2001) example of testing writing ability with an essay question which requires knowledge of history (when not all students are specialist enough on the topic) is a concern of content validity. Brown (2003) states that the content validity can be observed when the achievement is defined apparently. In sum, content validity is the evaluation of how well a test assesses the required representative language skills, structures, etc. with all relevant content. To exemplify, if a tester aims to measure the L2 speaking skills through conversations, asking test-takers to answer multiple choice questions conflicts with content validity. Yet, if the test demands the learner to speak in an authentic context, then the test is said to have content validity. Moreover, in order not to lack content

validity, a test also requires to cover all lesson objectives rather than just a few when there are more. Beside ensuring the content validity, the difficulty level of a task should also be reasonable with a feasible challenge.

Criterion-related validity, in the simplest term, indicates how precisely a test measures the outcome it was aimed to measure. Hughes (1989) expresses criterion-related validity as “how far results of the test agree with those provided by some independent and highly dependable assessment of the candidate’s ability” (p. 23). Akbari (2012) put it in the way as “if a test accurately measures a certain component or skill of the L2, it should closely correlate with other tests that measure the same component or skill” (p. 31). According to Weir (2005), criterion-related validity is at stake when test scores can be correlated with an older and longer test at the same time. In short, criterion-related validity emerges when a test can be validated via another independent assessment tool.

Criterion- related validity falls into two categories, which are concurrent validity and predictive validity. Concurrent validity can be obtained when the test and the criterion are implemented consecutively Hughes, 1989). Brown (2003) exemplifies this type of validity with a final exam and a proficiency exam administered afterwards. To him, the validity of the final test can be assured by a proficiency exam when the test-takers’ both scores display a high level of agreement. Another example comes from Hughes (1989). In an oral competency test, to be able to test all the course objectives may take around 45 minutes for each student, which is very impractical. Therefore, only ten minutes is allocated for each student. After the implementation of the test, some randomly chosen students are asked to perform the long run, which takes 45 minutes per student, and the scores on the full test and the ten-minute session test are compared to observe the concurrent validity.

The second category is predictive validity. As the name suggests, this type of validity involves how well a test foresees test-taker’s competency in future assessments. Brown (2003) illustrates the possible cases of predictive validity with placement tests and language aptitude tests. Another example, offered by Hughes (1989), can be a proficiency test which anticipates a student’s competency in withstanding a graduate course at a British university.

The third kind of validity is *construct validity*. Construct validity describes whether a test measures the ability that is supposed to be measured. Bachman and Palmer (1996) call

attention to the meaningful and appropriate interpretations of the test scores. To be able to do that, they ask “To what extent can we justify these interpretations?”, and come to the conclusion that test developers as well as test users need to yield sufficient justification for the interpretations made by the testers. To cite an instance, when a tester is assessing students’ communicative language ability in an oral interview, the ignorance of, let’s say, pronunciation and grammar, results in an invalid interpretation since an oral performance requires several factors including fluency, vocabulary use, sociolinguistic appropriateness as well as pronunciation and grammar (Brown, 2003).

Construct validity is rooted in scientific background, theories and hypotheses. It gains importance in validating large-scale standardized tests of proficiency such as TOEFL and International English Language Testing System (IELTS). Yet, although it is not the primary concern for language teachers, not only the test developers but also the test users should be aware that “construct validation is an ongoing process”, and that the interpretations never yield absolute validity (Bachman & Palmer, 1996, p. 22).

Consequential validity refers to the aftereffects of a particular test including test-related consequences such as accuracy in measuring the criteria to be assessed and its effects on the students, social consequences and washback. Bachman (1990) thinks that the use of language tests reflects the role of social policy, as well. Although tests are developed and utilized for various purposes like placement, diagnosis, etc., language tests are also implemented “to make decisions that have the societal implications such as the awarding of high school diplomas or the certification of language teachers on the basis of minimum competency tests” (Bachman, 1990, p. 280).

Face validity is the final type of validity, and concerns whether a test appears to measure what is attempted to measure. Harmer (2001) defines face validity as a test which looks “on the ‘face’ of it” as if it is valid” (p. 322). For Brown (2003), a test with face validity answers the question “Does the test, on the ‘face’ of it, appear from the learner’s perspective to test what is designed to test?” with a “yes” (p. 27). Hughes (1989) implies that a test without face validity may be difficult for test takers, teachers and education authorities to welcome, which may inappropriately manifest test-taker’s ability. Nonetheless, since face validity is not a scientific approach, it cannot be tested but is crucial in terms of validity. Brown (2003) indicated some factors that may increase the face validity. For example, if learners confront familiar, well-constructed tasks, without

vague items, and with sufficient time to perform, the test will probably entail high face validity. Besides, the instructions should be clear. Additionally, Burch and Kasper (2021) emphasize the importance of careful formulation of task instructions since they play a crucial role in eliciting the desired interactional competencies from test-takers. Clear and contextually rich instructions can pave the way to more authentic and effective assessments of speaking skills, thereby enhancing the test's face validity by aligning tasks with real-world communicative contexts.

2.7.4 Authenticity

In Cambridge Dictionary, authenticity is defined as “the quality of being real or true”. With regard to language learning, Bachman and Palmer describe authenticity as “the degree of correspondence of the characteristics of a given language test task to the features of a target language task” (Bachman & Palmer, 1996, p. 23), as displayed in the figure below.

Figure 2.6

Authenticity



Reference: Bachman & Palmer, 1996, p. 23.

In order to make generalizations about the test score interpretations, authenticity plays an important role. For one reason, “it relates the test task to the domain of generalization” (Bachman & Palmer, 1996, pp. 23-24). For Brown (2003), an authentic test also appears in the “real world” although Weir (1993) emphasizes the unavailability of “full authenticity of tasks or genuineness of texts” since it is not possible to completely replicate real life (p. 46). Another reason why authenticity gains importance is that it impacts the test takers’ perceptions, and as a result, their performance. Yet, authenticity does not merely involve topical content of the test. Likewise, the test item types ought to be relevant for the aim of the test (Bachman & Palmer, 1996). Brown (2003) suggests some ways to present authentic tasks, which are listed below:

- The language in the test is as natural as possible.

- Items are contextualized rather than isolated.
- Topics are meaningful (relevant, interesting) for the reader.
- Some thematic organization to items is provided, such as through a story line or episode.
- Tasks represent, or closely approximate, real-world tasks.

(Brown, 2003, p. 28)

2.7.5 Washback

Washback or backwash is defined by many scholars. Hughes (1989) describes it as “the effect of testing on teaching and learning” (p. 1). Brown (2003) explains washback as “the effects the tests have on instruction in terms of how students prepare for the test” (p. 28). Shohamy (2001) mentions washback as the effect of test, and is often used as impact, consequences or effect interchangeably. Hughes (1989) also declares that the effects of washback can be both positive and negative. If the test content and the testing techniques mismatch, it is highly probable to achieve harmful washback. For example, when the aim of the test is to measure writing skill, the multiple-choice test type is not acceptable because it does not help practice the skill itself. On the other hand, washback can be beneficial, too. Initially, it provides feedback to the learners, especially constructive feedback on their test performance may give them a sense of accomplishment, motivate them and foster learning. Furthermore, washback can help teachers and learners diagnose the challenging subsections in a test when the test scores display relatively low marks. In addition, students have the opportunity to discuss the feedback and evaluation they receive afterwards so that teachers can set up a constructive atmosphere where students can discuss and promote collaborative learning. That’s why Bachman and Palmer (1996) examine washback within the scope of “impact”. They believe that washback is a more complicated concept to be simplified as the effect of testing on teaching. Washback directly impacts not only the test-takers regarding the test experience, getting feedback and decisions made about them after getting their scores, but also the teachers. For instance, teachers are aware that their instruction is somehow affected by testing, or they may feel dissatisfied on the basis of the students’ test results. In addition to impact on individuals, washback, in a broader sense, can influence society and the education system with regard to the values and cultural norms (Bachman & Palmer, 1996).

Not many studies have attempted to analyze how to obtain beneficial washback, though (Weir, 2001). Hughes (1989) suggests some recommendations to attain beneficial backwash:

- Test the abilities whose development you want to encourage.
- Sample widely and unpredictably.
- Use direct testing.
- Make testing criterion-referenced.
- Base achievement tests on objectives.
- Ensure the test is known and understood by students and teachers.
- Where necessary, provide assistance to teachers.

(pp. 44-46)

To sum up, the five principles of a good language testing comprise practicality, reliability, validity, authenticity and washback, all of which are necessary in the evaluation of an existing test as well as designing a brand new one since they affect a considerable number of factors which are at play such as teachers, learners, methodologies, and learning.

2.8 Characteristics of Test Tasks

Most language teachers are familiar with the importance of the task types in language tests and think of the best way of measuring a particular language ability. How language teachers test the learners' ability influences the students' performance on the tests and, as a result, the characteristics of the information their tests yield. Weir (1993), for example, underlines the importance of "performance conditions", "operations" and criteria for assessment since it will lead to the "quality of output". Bachman and Palmer (1996) put emphasis on the specific characteristics of the test tasks and the appropriateness of these tasks for particular groups of test takers. Hence, they draw a framework of language task characteristics which involve:

1. describing target language use (TLU) tasks as a basis for designing language test tasks,
2. describing different test tasks in order to insure their comparability, and as a means for assessing reliability, and

3. comparing the characteristics of TLU and test tasks to assess authenticity.

(Bachman & Palmer, 1996, p. 47)

The table below outlines the test task characteristics, which will be discussed presently.

Table 2.1

Task Characteristics

Task characteristics
Characteristics of the setting
<i>Physical characteristics</i>
<i>Participants</i>
<i>Time of task</i>
Characteristics of the test rubrics
<i>Instructions</i>
Language (native, target)
Channel (aural, visual)
Specification of procedures and tasks
<i>Structure</i>
Number of parts/tasks
Salience of parts/tasks
Sequence of parts/tasks
Relative importance of parts/tasks
Number of tasks/items per part
<i>Time allotment</i>
<i>Scoring method</i>
Criteria for correctness
Procedures for scoring the response
Explicitness of criteria and procedures
Characteristics of the input
<i>Format</i>
Channel (aural, visual)
Form (language, non-language, both)
Language (native, target, both)
Length
Type (item, prompt)
Degree of speededness

Table 2.1 (continued)

Task Characteristics

Vehicle ('live', 'reproduced', both)
<i>Language of input</i>
Language characteristics
Organizational characteristics
Grammatical (vocabulary, syntax, phonology, graphology)
Textual (cohesion, rhetorical/conversational organization)
Pragmatic characteristics
Functional (ideational, manipulative, heuristic, imaginative)
Sociolinguistic (dialect/variety, register, naturalness, cultural references, and figurative language)
Topical characteristics
Characteristics of the expected response
<i>Format</i>
Channel (aural, visual)
Form (language, non-language, both)
Language (native, target, both)
Length
Type (selected, limited production, extended production)
Degree of speededness
<i>Language of expected response</i>
Language characteristics
Organizational characteristics
Grammatical (vocabulary, syntax, phonology, graphology)
Textual (cohesion, rhetorical/conversational organization)
Pragmatic characteristics
Functional (ideational, manipulative, heuristic, imaginative)
Sociolinguistic (dialect/variety, register, naturalness, cultural references, and figurative language)
Topical characteristics
Relationship between input and response
<i>Reactivity</i> (reciprocal, non-reciprocal, adaptive)
<i>Scope of relationship</i> (broad, narrow)
<i>Directness of relationship</i> (direct, indirect)

Reference: Bachman & Palmer, 1996, pp. 49-50.

2.8.1 Characteristics of the setting

Characteristics of the setting is about the physical circumstances in which language use and testing actualize. These characteristics are physical setting, participants and time of task.

Physical setting comprises the location, the physical conditions such as temperature, humidity, lighting, noise level along with seating arrangements. Bachman and Palmer (1996) believe that in a place where a test is administered, materials and equipment are arranged accordingly such as pencils, paper, audio-visual equipment, and so forth.

Participants refer to the people who take part during the task including both the test takers and the testers. The relationship and the familiarity of these people shape the characteristics of the setting. For example, in some test tasks, test takers engage with each other to be able to negotiate the meaning, whereas in others like a reading task the roles of participants may change, and they do not need to engage with one another.

As is suggested by the name, time of task corresponds with the time when the test is administered. The time of the test may have an influence on students' performances, either in a positive or negative way. For example, the time of the test may hinder the learners' ability to perform at their best and can cause fatigue. Conversely, depending on the test time, they may feel fresh.

2.8.2 Characteristics of the test rubrics

Test rubrics guide the test takers and the test administrators how to proceed to be able to accomplish the task. Thus, it is highly essential to design the test rubrics as explicit and clear as possible. The characteristics of the test rubrics comprise instructions, structure, time allotment and scoring method.

Explicit instructions are essential since the learners' test performances depend on how much they have understood the procedure during the test. Instructions also provide information for test takers about how to be scored and how the results will be utilized. The test administrators also pay attention to the language preferences while giving the instructions. They can be either in the native language or the target language or, in some circumstances, both. Also, the channel, which can be aural, visual or both should be taken into consideration. Lastly, specification of procedures and tasks need to be clearly

specified for the test takers. For example, whether the instructions are lengthy or brief, whether they are presented with or without examples or whether the particular parts are linked to another part or not should be considered properly and explained explicitly.

The structure of the test is about the organization of the test. Specifically, it includes the composition of parts in the test and the presentation of these parts to the test takers. First of all, the number of parts/tasks in the test should be specified. Also, the salience of parts/tasks is important. That is, the test takers are easily able to figure out the different parts of the test. To be more specific, whether there is a single task or separate tasks can be distinguished without a hitch. The sequence of the parts/tasks is another important characteristic of instructions since it displays whether students respond to the test tasks in a fixed order or not. Besides, the instructions need to specify the importance of the test tasks if there is any difference. Finally, the number of the tasks in the test together with the items in each part should be pointed out in the instructions.

The third characteristics of the test rubric is the time allotment, referring to the duration of the entire test and of the separate parts. The test can be designed as a *speeded*, which means not all participants of the test are able to finish all tasks or a *power test*, in which each learner has ample time to go through every task.

Scoring method is the final characteristics of the test rubric, which involves the evaluation of each test taker's performance. To this end, objective scoring criteria and accurate procedures for scoring the responses such as tests' being scored in a particular sequence or all performances' being rated by the same rater, etc. gain importance. Moreover, how much the test takers have been notified about the scoring criteria and the procedures need to be explicit enough.

In conclusion, test rubrics have some characteristics including instructions, structure, time allotment and scoring method to be able to assemble more appropriate language test tasks that fit the purpose.

2.8.3 Characteristics of the input

Input is the material given in a test task, and is addressed through format and language characteristics. Format is the way of the presentation of the input. Therefore, channel, form, language, length, type, degree of speededness and vehicle has to be reviewed. First, the channel could be in aural, visual or in both versions. Second, form can be given via

language or with non-language such as pictures and gestures, or both. Likewise, the language of input can be either in native or target language, or both. Next, length means whether the input includes single words, phrases, sentences, or paragraphs and extended discourse. In order for the test takers to interpret less, it is much better to avoid extended discourse. Fourth, type of input can be separated as item and prompt. A test item aims to elicit response as in multiple-choice or completion questions. A prompt, however, is the input given by the tester to elicit longer production responses as in a writing task, for instance. Suppose that students are asked to describe their summer holiday, which is also a prompt. Fifth, degree of speededness concerns the rate of the test taker's information process in the input. Finally, vehicle refers to how the input is delivered. It can be either live, reproduced or both. Bachman and Palmer (1996) exemplify this with a use of lecture in a listening comprehension test, in which the lecture can be given either live through audio or reproduced through videotape.

The second characteristic is the language of input, which comprises language characteristics and topical characteristics. Language characteristics include organizational characteristics including grammatical characteristics like vocabulary, syntax, phonology, etc. and textual characteristics like cohesion, rhetorical or conversational organization. Topical characteristics are the type of personal, cultural, academic, or technical information that is found in the input (Bachman & Palmer, 1996).

2.8.4 Characteristics of the expected response

Similar to the characteristics of the input, characteristics of the expected response can be examined in terms of format and language. Bachman and Palmer (1996) define format as "the way in which the response is produced" (p. 54). Test specialists divide test takers' responses into two categories: selected response, in which students select an answer among other options, and constructed response, in which the learner needs to produce a response. Yet, Bachman and Palmer (1996) split constructed responses into two depending on the length of the production. They call limited production response for the single word or phrases or a single sentence or utterance, whereas extended production response consists of a range of sentences and utterances that comprise an oral or written composition. Degree of speededness is the amount of time a test taker has to give a response. Like characteristics of input, degree of speededness concerns the rate of the test

taker's response. That is, if the score of the test necessitates a quick response, the task is called speeded.

Second characteristic of expected response is related to the language. As mentioned in the characteristics of the input, the same characteristics exist, which are channel, language, length, genre, content, and organization (Bachman & Palmer, 1996). To illustrate, responses can be in either spoken or written channel, using the same language as the input or a different one. The length of the answer may be a single word or an extended written /oral discourse, and the genre might vary from informal conversation to academic essay. Similarly, the content of the response must be compatible with task requirements, and its organization should reflect logical structuring suitable to the task type such as using chronological order in a narration.

2.8.5 Relationship between input and response

Bachman and Palmer (1996) also describe the relationship between input and response in terms of reactivity, scope of relationship and directness of the relationship.

Reactivity is the direct effect of input or response to the following input and response. It is explained under the three test tasks, which are reciprocal, non-reciprocal and adaptive. In a reciprocal task, there is an interaction between the test taker and the interlocutor. Also, explicit (verbal) or implicit (physical) feedback is given by the interlocutor subsequently. Thus, the test taker's language use is affected by communicative exchange, as in a face-to-face oral interview. In non-reciprocal tasks, no feedback or interaction between the test taker or the test administrators exists. A typical example can be a reading task, in which the test taker's response to the question does not affect the following test task. Adaptive tasks are the tasks in which the subsequent task is determined by the test taker's performance on the previous task. The first task presents a moderate level of difficulty, and the subsequent task becomes comparatively more challenging if the first task is accurate. Or if the test taker does not respond to the question accurately, a slightly easier question is asked. Although adaptive tasks seem similar to the reciprocal tasks, in most adaptive tasks, no feedback is provided. However, they involve interaction to some extent as the test taker's response characterizes the following question. The differences of these tasks are given in the table below.

Table 2.2*Distinguishing features of reactivity between input and response*

Relationship between input and response	Feedback: knowledge of relevance or correctness	Interaction: response affects subsequent input
Reciprocal	+ (present)	+ (present)
Adaptive	- (absent)	- (absent)
Non-reciprocal	- (absent)	- (absent)

Reference: Bachman & Palmer, 1996, p. 56.

Scope of relationship refers to the amount of input that a test taker has to process to be able to answer the question. Scope can be either broad or narrow. Broad scope tasks involve a lot of input to be processed as in the example of getting the gist of a listening or a reading task, which the test taker looks for the main idea without heeding the specific details. Narrow scope tasks, on the other hand, involve only a limited amount of input to be processed. Reading for specific details or scanning a text to find specific information are examples of narrow scope tasks.

Directness of the relationship is whether the test taker needs to consult the information in the input or in his own topical knowledge to be able to respond to the question. It can be either direct or indirect. In a direct relationship, the response depends on the input. A picture description task in a speaking test where the test takers have to rely on the content of the picture is an example of direct relationship. In indirect tests, the response does not necessarily depend upon the input. For example, a speaking test where the test taker expresses his/her opinion about a recent event would be an example to a relatively indirect test.

2.9 Characteristics of Test-Takers

Many researchers agree that several individual characteristics, the vast majority of which are not predicted easily such as fatigue or are not easily measurable such as motivation and introversion, influence the learner's rate of learning and language use (e.g., Bachman & Palmer, 1996; Harmer, 2001; Lightbown & Spada, 2021). Yet, Bachman and Palmer (1996) try to compile certain characteristics whose impacts on learner's language test performance are recognized better, which are personal characteristics, topical knowledge, affective schemata and language ability.

2.9.1 Personal characteristics

Personal characteristics are the individual qualities, and may have an impact on language learners' test performances. For Bachman and Palmer (1996), these characteristics are "age, sex, nationality, resident status, native language, level and type of general education, and type and amount of preparation or prior experience with a given test" (p. 65). It is important to consider these characteristics while developing a test to maximize the usefulness of it.

2.9.2 Topical knowledge

Bachman and Palmer (1996) define topical knowledge as "knowledge structures in long-term memory" (p. 65). Topical knowledge, in other words, is the learner's prior knowledge of a specific topic. Topical knowledge allows learners to use language with regard to the world they live in. Certain test tasks may be easier for learners who already have that cultural or topical knowledge compared to the ones who do not (Bachman & Palmer, 1996). To illustrate, it might be expected for a business student to perform a better speaking task on marketing. Similarly, when a reading text that mainly includes particular cultural information may be troublesome for a student who does not belong to that culture. Thus, it can be said that topical knowledge has an impact on test takers' speaking performance.

2.9.3 Affective schemata

Affective schemata are the emotional connections to topical knowledge, which are, to a large extent, based upon past experiences. These affective schemata determine the principles for language users to access the language task, and either in a facilitative or a restrictive way. That is to say, affective schemata influence the test taker's response to the task in a positive or negative way. Experiencing strong, uninhibited emotions may restrict their language ability to use certain knowledge and metacognitive strategies. On the other hand, emotional responses can be quite facilitatory. Bachman and Palmer (1996) illustrate this facilitative emotional status with a speaking exam, in which individuals with positive feelings about interaction with others tend to perform better in face-to-face interviews. After all, some language test designers draw attention to the design of test tasks since they take heed of the tasks which can elicit test takers' best performance by minimizing the threads that some students may feel because of the type of test tasks, etc.

2.9.4 Language skills

According to Bachman (1990), language skills are composed of two elements: *language knowledge* and *strategic competence*. *Language knowledge* refers to the information stored in the memory and is ready for use with the help of metacognitive strategies. Language knowledge also involves two components: *organizational knowledge* and *pragmatic knowledge*.

Organizational knowledge mainly concerns how utterances or sentences and texts are arranged to be able to produce and understand grammatically accurate utterances/sentences and texts in spoken and written forms. Organizational knowledge can also be divided as grammatical knowledge and textual knowledge. Grammatical knowledge consists of vocabulary, syntax, phonology and graphology, which are necessary to produce and comprehend accurate sentences. Textual knowledge entails producing and understanding units of spoken or written language. Textual knowledge has two main concepts: knowledge of cohesion and knowledge of rhetorical or conversational organization. The former area is associated with the relationships among sentences in a written text or utterances/sentences in spoken language whereas the latter focuses on the organization development of written and oral tasks.

Pragmatic knowledge helps us develop discourse by interpreting the meanings of texts, intentions of language users and characteristics of the language use. Pragmatic knowledge has two areas, as well: functional knowledge and sociolinguistic knowledge. Functional knowledge involves the association between the utterances/sentences or texts with the language users' intentions. Functional knowledge is classified in four categories: in terms of language functions: ideational, manipulative, instrumental and imaginative. Sociolinguistic knowledge is more related to the setting in which the language use emerges. Thus, it involves dialects, varieties, register, natural or idiomatic expressions and figures of speech.

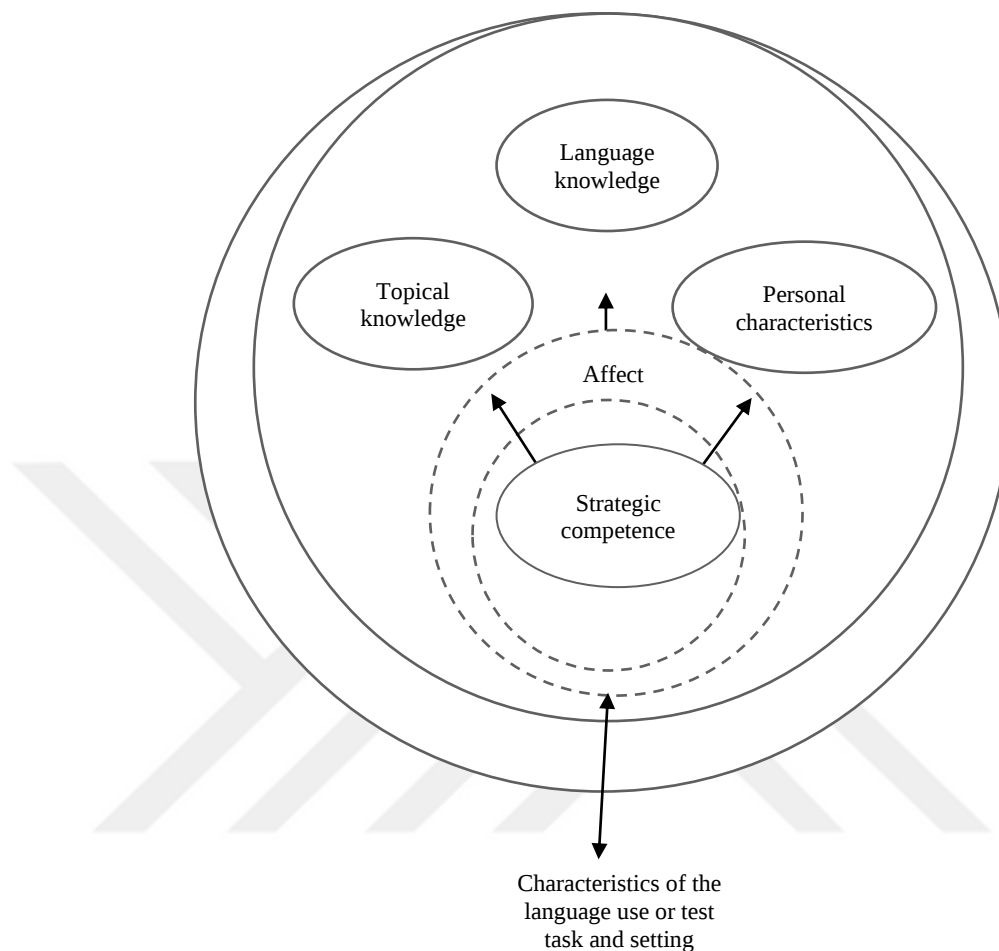
Strategic competence is a higher order process which requires the ability to use metacognitive strategies. Bachman and Palmer (1996) underline the importance of strategic competence in both designing and developing interactive test tasks and also evaluating these test tasks. Strategic, or metacognitive, competence incorporates goal setting, assessment and planning.

While designing a test, it is important to develop them which are able to measure a larger scope of elements of language ability. Goal setting includes identifying the test tasks, choosing one or more test tasks, where given a choice and making a decision on completing the task or not. Assessment involves “assessing the characteristics of the language use or test tasks” (p. 71), assessing the individual’s own topical knowledge and language knowledge” and assessing the correctness or appropriateness of the response to the test task” (Bachman & Palmer, 1996, p.73). Planning is the use of language knowledge, topical knowledge and affective schemata to be able to execute the task in a successful way. Therefore, planning includes choosing certain elements from topical knowledge and language knowledge such as words, concepts and structures. Planning also necessitates specifying plans prior to the response to the task such as interpretation and utterance as well as choosing a plan to implement the task.

The figure below is an example of some components of language use and language test performance (Bachman & Palmer, 1996).

Figure 2.7

Some components of language use and language test performance



Reference: Bachman & Palmer, 1996

2.10 Speaking Skills

In the realm of language proficiency, there exist four core language skills, namely; reading, writing, listening and speaking. These skills are fundamental components that underpin effective communication and language competence. However, among all of them, the attainment of proficiency in speaking is widely regarded as the most challenging skill for language learners. Horwitz, et. al. (1986) state speaking as one of the most anxiety-provoking aspects of language learning. Hedge (2001) says that speaking skill in a competent way is a complex task. This complexity arises from the multifaceted nature of oral communication, which encompasses pronunciation, fluency, grammar, vocabulary, cultural nuances, and the ability to engage spontaneous conversations (e.g., Harris, 1990). Yule (2010) underlines the importance of three components of

communicative competence: grammatical, competence, linguistic competence and strategic competence. These three components are essential for language learners to be able to speak accurately, appropriately and flexibly. Harmer (2001) underlines two features to be able to speak fluently. The first one is the language features encompassing connected speech, expressive devices, lexis and grammar, and negotiation language. Mental/social processing is the second fundamental feature of oral communication ‘on the spot’. It encompasses language processing, interacting with others, and (on-the-spot) information processing. Hedge, also, (2001) draws attention to the components that are entailed in achieving speaking competence in the use of English language. These are distinguishing types of speaking situations, making oneself understood, and managing interaction. Since speaking competently is a demanding situation for language learners, the skill is regarded as one of the most challenging ones.

Bygate (1987), on the other hand, emphasizes the difference between ‘knowledge’ and ‘skill’ with an analogy of a car driver. While knowing the names of the controls, their places and their functions is the ‘knowledge’, using these controls to drive the car without trouble is the ‘skill’. Therefore, both knowledge and skills are essential for a language learner to achieve competency. He also underpins the differences between productive skills not only in its grammatical, lexical and discourse patterns but also ‘processing conditions’. He illustrates the effect of time pressure on the speakers in two ways. First, speakers utilize certain devices to facilitate production and gain time. Speakers can facilitate their speech by simplifying the structure, by ellipsis, by using formulaic language and by the use of fillers and hesitation devices (Bygate, 1987, p. 15). Second, they generally need to compensate in problematic situations since they have less time to plan, organize and execute their response. Hence, in oral communication, alterations are acceptable and tolerated to compensate for the difficulties via adjustments such as hesitations, false starts, self-corrections; syntactic features like ellipsis; repetition through expansion or reduction; and formulaic expressions.

All in all, speaking competently can be a demanding skill for language learners, and, thus, regarded as challenging. Therefore, learners use facilitation and compensation strategies and employ negotiation skills for oral production in a competent way.

2.11 Speaking and Language Acquisition Theories

Speaking can be associated with specific theories of second language acquisition (SLA). To start with, it would be quite valid to interpret the speaking skill with Vygotsky's Sociocultural Theory (ST) (1978), which also holds a very important place in SLA. According to Richards and Rogers (2014), the term sociocultural means that learning takes place in a “particular social setting (e.g., a classroom), in which there is interaction between people (teachers and students), objects (texts, books, images), and culturally organized activities (instructional acts and sequences)” (p. 28). In his theory, Vygotsky emphasizes that speaking is seen as deeply entwined with social interactions and cultural contexts. That is, the theory highlights the role of social interaction along with the cultural context. As cited by Thornbury and Slade (2006), this mindset of learning “focuses primarily on the social and cultural processes that contribute to the development of higher order cognitive functioning” (p. 207). In the opinion of Swain (2005), ST “puts language production in a ‘star role’ so to speak” (p. 480). To be brief, language development occurs through social interactions, specifically, with more knowledgeable individuals, which has also given rise to a new term, scaffolding, in which more educated or knowledgeable entities like teachers and peers who can furnish support and guidance to learners. Scaffolding plays an important role since it facilitates the learners’ acquisition of oral skills through structured interactions and participation in meaningful communicative activities. Thus, speaking proficiency is not solely the result of individual effort but is firmly grounded in social and cultural practices within the sociocultural context. Zone of proximal development (ZPD) is another directly related concept to ST. According to Lantolf (2005), “The ZPD is a projection of a person’s developmental future in the sense that what one can do in cooperation with others today one can do alone tomorrow” (p. 336). ZPD refers to the hypothetical area in which the learner’s abilities take place. It also represents the difference between what a learner can do without help and what s/he can do with guidance and support. The theory suggests the expansion of ZPD via social interaction “with peers, with more knowledgeable others, and through one’s own resources” (Bailey, 2020, p. 53). All things considered, when speaking assessment is taken into account, test tasks involving collaborative activities, projects and/or group discussions can encourage language learners to be active in language acquisition process in social settings.

Second, Long (1981) put forward Interactionist Hypothesis (IH), which is closely related to the oral communication skills. Similar to ST developed by Vygotsky, this hypothesis highlights the value of interaction and social context. Namely, meaningful interaction in the target language enhances language acquisition. Learners' being involved in authentic conversations and communicative exchanges with native or more proficient speakers promotes the acquisition of speaking skills. Furthermore, Long (1981) emphasizes that through interactive communication, language learners can receive feedback and negotiate meaning, which ultimately leads to improved speaking abilities. Similarly, Rost (2016) mentions three key concepts of IH. The first one is "learners are the ones doing the learning" (p. 137). Second, "learning requires negotiation for meaning", and finally, "feedback is necessary for learning" (p. 137). In essence, regarding speaking skills development, the IH suggests that meaningful interaction plays a fundamental role in SLA. In speaking assessment, tasks with a purpose of real or simulated communication with peers can reflect this approach by testing learners' ability to negotiate meaning, participate in authentic conversational exchanges and respond to feedback.

Another exemplary instance that certainly springs to mind is the Affective Filter Hypothesis (AFH) proposed by Steven Krashen (1982), who is widely recognized as one of the pioneers in the field. According to his hypothesis, an individual's emotional state, such as anxiety and motivation, may have a significant impact on language acquisition. Namely, this theory posits that a low affective filter, characterized by low anxiety and high motivation, facilitates the acquisition and production of oral language. Conversely, a high affective filter, emerging from anxiety or other unfavorable emotions, can restrict language production and, thus, hinder oral proficiency. As a result, creating a supportive and low-anxiety learning environment is crucial for learners to feel comfortable engaging in spoken communication and enhancing their speaking skills. Regarding speaking assessment, this theory can inform the test designers and the test implementors that creating a supportive testing setting can minimize the anxiety level of testers and provides a less stressful testing environment for the learners so that a more comfortable and confident environment can be fostered during the testing process.

Beside ST, IH and AFH, Swain's Comprehensible Output Hypothesis (COH) (1985) can also be associated with oral skills since "the processes involved in producing language can be quite different than those involved in comprehending language" (Swain, 2005, p.

471). In the context of speaking skill development, Swain's COH gives prominence to producing language, particularly through language output activities such as writing and speaking. To achieve this, Swain (2005) gives prominence to three beneficial functions of output. First one is *noticing/triggering*. According to COH, when learners actively engage in speaking tasks, they start to benefit from opportunities to produce language output by noticing gaps in their linguistic knowledge and seeking further clarification or correction. Second beneficial function is *hypothesis testing*. Swain (2005) calls it as “trial run” where learners test their use of language (p. 476). By trying to express themselves verbally, learners are encouraged to utilize their existing language skills and generate new language structures. This process of "pushed output" fosters language learning by motivating learners to produce language that is beyond their current level of proficiency. The *metalinguistic (or reflective) function* is the third function which focuses on “using language to reflect on language produced by others or the self mediates second language learning” (Swain, 2005, p. 478). Hence, according to COH, the enhancement of speaking skills is fostered through active language production and feedback attainment on their output. In terms of speaking assessment, test tasks based on this approach need learner production including presentations, role-plays and discussions.

Moreover, it is essential to consider DeKeyser's (2007) Skill Acquisition Theory (SAT), which provides a useful lens to understand the development of speaking proficiency in SLA. DeKeyser (2007) points that language learning involves progressing from declarative knowledge, which is explicit understanding of rules, to procedural knowledge, where learners start to use the target language automatically and fluently in real-time communication. In speaking, this transition is especially vital because effective oral communication requires quick retrieval and use of language structures without conscious rule application. Learners at the declarative stage may hesitate, self-correct excessively, or struggle with fluency during speaking tasks. Thus, DeKeyser (2007) emphasizes the concept of proceduralization, which enables learners to produce language more spontaneously and confidently, particularly under the time constraints and social pressure of speaking exams. To achieve this, repeated and meaningful speaking practice gains importance so that learners can move beyond rule-memorization to more automatic and fluent language use.

In essence, by integrating SLA theories into the assessment of speaking skills, educators gain a deeper understanding of how learners develop and apply language proficiency. Furthermore, test designers can take into account the priorities and characteristics emphasized by these SLA theories when preparing test tasks and the test environment, and likewise test administrators can take into account the priorities and characteristics emphasized by these theories during administration, which leads to more effective and comprehensive evaluation methods tailored to language learners' needs.

2.12 Testing Speaking Skills

Testing speaking skills is a critical component of language assessment, as it provides valuable insights into an individual's ability to communicate effectively in real-life situations. Unlike other areas of language proficiency, such as reading, listening or writing, speaking is highly dynamic and interactive, which encompasses both the reception and the production of language. As such, the process of testing speaking typically involves designing tasks or prompts that require test-takers to express themselves orally. These tasks can range from structured exercises, such as reading aloud or describing images, to more open-ended activities like discussions or role-play scenarios. The assessment criteria often encompass pronunciation, vocabulary usage, grammar accuracy, fluency, and overall coherence in communication. To accurately assess speaking proficiency, it is essential to consider both the linguistic aspects of speech like grammar and vocabulary as well as the overall communicative competences such as pragmatics and discourse management, taking into account the speaker's ability to comprehend and respond appropriately to different conversational contexts and situations. Effective assessment methods typically involve a combination of structured tasks supplying a clear framework for assessment, and spontaneous interactions that mimic real-life conversations, and as a result, allow for a more authentic measure of proficiency.

One of the fundamental distinctions in the assessing speaking is between *direct* and *indirect* assessment approaches (Louma, 2004). *Direct assessment* asks learners to produce spoken language and involves observing learners while they perform oral tasks, such as interviews, role-plays, group discussions or presentations in real-time; thereby, it offers a more authentic representation of students' communicative ability. Since it directly measures the target skill, direct assessment of speaking is favored more (Louma,

2004). *Indirect assessment*, in contrast, measures speaking ability through tasks that do not involve actual speaking. It evaluates the underlying components of speaking such as grammar and/or vocabulary without spoken output, which loses its validity of real-world speaking performance.

In addition, while assessing speaking, Brown (2003) defines 16 micro- and macroskills of oral production among which the testers choose as the objective(s) of the assessment task(s). Microskills refer to the production of small chunks of language including phonemes, morphemes, words and collocations. On the other hand, macroskills focus on larger communicative units such as fluency, discourse, function, style, cohesion and strategy use (Brown, 2003). These macroskills are important in assessing a speaker's ability to use language appropriately across different social contexts. The table below (Table 2.3) lists these micro- and macroskills of spoken production.

Table 2.3

Micro- and macroskills of oral production

Microskills 1. Produce differences among English phonemes and allophonic variants. 2. Produce chunks of language of different lengths. 3. Produce English stress patterns, words in stressed and unstressed positions, rhythmic structure, and intonation contours. 4. Produced reduced forms of words and phrases. 5. Use an adequate number of lexical units (words) to accomplish pragmatic purposes. 6. Produce fluent speech at different rates of delivery. 7. Monitor one's own oral production and use various strategic devices – pauses, fillers, self-corrections, backtracking – to enhance the clarity of the message. 8. Use grammatical word classes (nouns, verbs, etc.), systems (e.g., tense, agreement, pluralization), word order, patterns, rules, and elliptical forms. 9. Produce speech in natural constituents: in appropriate phrases, pause groups, breath groups, and sentence constituents. 10. Express a particular meaning in different grammatical forms. 11. Use cohesive devices in spoken discourse. Macroskills 12. Appropriately accomplish communicative functions according to situations, participants, and goals.
--

Table 2.3 (continued)

Micro- and macroskills of oral production

13. Use appropriate styles, register, implicature, redundancies, pragmatic conventions, conversation rules, floor-keeping and -yielding, interrupting, and other sociolinguistic features in face-to-face conversations.
14. Convey links and connections between events and communicate such relations as focal and peripheral ideas, events and feelings, new information and given information, generalization and exemplification.
15. Convey facial features, kinesics, body language, and other nonverbal cues along with verbal language.
16. Develop and use a battery of speaking strategies, such as emphasizing key words, rephrasing, providing a context for interpreting the meaning of words, appealing for help, and accurately assessing how well your interlocutor is understanding you.

Reference: Brown, 2003, pp. 142-143.

2.12.1 Types of oral tests

There are various question types commonly used in testing speaking skills, each serving different assessment objectives and focusing on distinct aspects of oral communication proficiency. The choice of test type should be aligned with the assessment's purpose and the communicative demands of the target language use. In this respect, common test formats include interviews, role-plays, picture description tasks, and paired or group discussions (Luoma, 2004). Yet, depending on the levels of complexity and the range of communication skills required, Brown (2003) divides speaking tests into five categories:

- imitative,
- intensive,
- responsive,
- interactive and
- extensive.

2.12.1.1 Imitative tasks

As the name suggests, in imitative speaking, the test-taker repeats a word, or a phrase, or even sometimes a sentence. Though it was considered outdated, Brown (2003) suggests it has gained interest in recent years since too much attention is given to fluency which has outweighed the place of accuracy in speaking. Therefore, a special attention has been given to pronunciation, particularly suprasegmentals, so that learners can be more intelligible.

One of the best-known examples of imitative speaking is word repetition. In this type of task, the test-taker hears a word, a phrase or a sentence and dictates the stimulus. Reading aloud the written stimulus is another variation of this kind of task. An additional example is the PhonePass Test, which is popular in many countries. The test is comprised of five sections, and the first and second sections are mainly an illustration of imitative speaking, in which test-takers read aloud some sentences in the first section and repeat the sentence dictated over the phone in the second section.

In short, imitative speaking tests can be useful when the purpose is to test the test-takers' pronunciation and intonation competencies.

2.12.1.2 *Intensive tasks*

In intensive speaking tests, the test-takers are asked to produce short discourse structures to demonstrate their linguistic skills at their capacity level. The tasks are generally cued in that the test takers' response possibilities are foreshortened. As mentioned above, Sections C and D in PhonePass, which has five sections in total, are examples of intensive speaking in which test-takers answer certain questions with one or two words and link word groups they hear in an accurately ordered sentence, respectively.

Another illustrious example of intensive testing is directed response tasks. In this mechanical task, the tester aims to receive a linguistic form or sentence transformation, and expect an accurate grammatical output. Read-aloud tasks are the other well-known intensive speaking tasks, which can vary from one sentence to a paragraph. In this controlled task, the results of the task indicate the test-takers' phonological capability along with the fluency. Intensive reading can be advantageous in terms of easy and reliable scoring and practicality; however, it also has some drawbacks. Initially, it is not authentic as not many people read something aloud in real life. In addition, it also requires certain oral abilities rather than communicative and pragmatic abilities. Thus, Brown (2003) advises the test administrators to employ this type of testing cautiously.

Another technique is sentence/dialogue completion tasks. In this type of task, test-takers read a dialogue in which some lines are omitted. A tape or the test administrator reads one part and the learner responds relationally. Likewise, learners can answer the questions or generate appropriate questions by looking at the written stimulus. Moderate control of the student production can be regarded as an advantage. Yet, being inauthentic and relying

too much on reading comprehension and transfer ability, that is test-takers must understand the written stimulus and transfer it to speech production, are the main drawbacks.

Picture-cued tasks are also among the most popular intensive speaking tasks. Test-takers look at some visual stimulus to talk, which can be as simple as solely eliciting a word, or more demanding as producing a story by looking at a series of pictures. The more the difficult the tasks become, the more challenging the process is for both the test-taker and the examiner. Test-takers need to produce relevant linguistic stretches while fulfilling the requirements of the task. Similarly, the examiner needs to be cautious while trying to correctly understand the examinees and determining the incorrect uses.

Final type of intensive speaking task is translation. Though it seems old-fashioned, translation of limited stretches of discourse can be an interpreter of learners' oral production skills.

2.12.1.3 *Responsive tasks*

While responsive tasks share similarities with intensive tasks, the key distinction lies in the interactive nature of the former one. That is, responsive tasks require an interaction between an interlocutor and the test-taker. Therefore, the test-taker is more creative in the sense of oral production.

One common example is question-and-answer tasks. In this technique, the interviewer can ask one or two questions, or s/he can ask a series of questions with prompts. The test administrator needs to understand the purpose behind each question. Therefore, the questions can be classified as display questions, with predetermined correct responses, or referential questions, allowing the test taker more freedom to generate meaningful language in response. Examples of potential question-and-answer tasks include open-ended questions like "What are the key skills necessary for language acquisition?" or "What motivated your choice of this department?" On the other hand, it is also feasible to reverse the role of questioning to evaluate the test-taker's proficiency in formulating questions, such as "Ask me about my family, profession, or hobbies."

Giving instructions and directions is another technique for oral assessment, which requires test-takers producing relatively longer stretches of discourse. In this technique, the tester states a problem, and the test-taker is expected to respond as in describing how

to create a dish, how to put a bookshelf together, giving directions to a foreigner, etc. Comprehensibility is paramount in the assessment, followed by grammar and discourse.

Paraphrasing is another type of responsive test task, which requires the test-taker to read or hear a limited number of sentences and paraphrase them. The task may involve either paraphrasing a brief story or delivering a phone message, with the latter being more authentic since it requires the test-taker to convey the message to another person. One benefit of this type of task is its ability to elicit brief spoken segments and lessen the burden of input/output ratio during a conversation. Nonetheless, the listening ability of the test-taker comes into play, which may affect the scoring, as well.

2.12.1.4 *Interactive tasks*

When compared to previously mentioned test tasks, interactive speaking tasks encompass extended exchanges of discourse. Interviews, role-plays, discussions and games are the basic examples to this type of assessment.

Oral interviews may be the most familiar examples that springs to mind. During an oral interview, an examiner, occasionally with a partner, engages in a direct face-to-face interaction with an examinee, or sometimes two, following a set of predetermined questions and guidelines. Based on the intended goals and context, the duration of the interviews can change from five minutes to forty-five minutes. Various components may be taken into consideration while scoring including accuracy, fluency, language use, task accomplishment, etc. According to Brown (2003), a successful interview must have some stages: warm-up, level check, probe and wind-down. In the first stage, the purpose is to calm down the examinee with a preliminary small chat. This phase does not take long, and the performance is not graded. In the level check stage, a set of predetermined questions are probed, and the examinee is supposed to use certain predicted or expected language structures. The test-taker's responses may vary depending on his/her level and the complexity of the questions. In the probe stage, however, the test-takers are challenged with more difficult questions to go beyond their level to observe the examinee's utmost proficiency. The final stage, wind-down, is a closure stage where the test administrator asks some simple questions to cool down the test-taker and gives some details about when and how to get the test results. As in the initial stage, this final stage is not included in scoring.

Role-play is another well-known way of interactive testing. It can promote the test-taker to use his/her creativity depending on the amount of control of the task. Role-playing is an efficient technique to elicit the targeted discourse such as conversation of a waiter and a customer at a restaurant, or a doctor and a patient at a hospital, which also displays real-world pragmatics.

An additional example to interactive test tasks can be discussions and conversations. Brown (2003) pinpoints the difficulties of determining and assessing discussions and conversations, yet he also suggests using them as an informal technique to assess learners' ability. Therefore, it may be better to use them in the classroom because discussions and conversations can also be regarded as authentic. In addition, the assessing scoring and checklists should be meticulously designed so that they can serve for the purpose.

Games, by the same token, are deemed as informal types of assessment, and can be various in number. Although these types of tasks are seen as untraditional, the correct, reasonable and practical set of criteria can make it a way of informal assessment due to its formative nature.

All in all, interactive speaking test tasks require longer stretches of discourse and can be utilized in not only summative but also formative evaluation.

2.12.1.5 *Extensive tasks*

Similar to interactive tasks, extensive speaking tasks also require longer stretches of discourse but less interaction as in oral presentations, picture-cued story-telling, retelling a story/news event and translation of extended prose.

Oral presentations are commonly found in academic and professional contexts, involving the delivery of reports, the introduction of marketing strategies or new sales concepts, and the presentation of newly developed products. Yet, it is imperative to prepare meticulous assessment criteria. Brown (2003) suggests a checklist or grid for the assessment of oral presentations including several components such as content and delivery.

Picture-cued storytelling is an additional method for eliciting oral production. While previously categorized as an intensive task, it can also be viewed as an extensive task when a single picture or a sequence of pictures is utilized to generate a more extensive narrative. Similar to earlier tasks mentioned, the assessment criteria must be unambiguous and the specific aspects to be assessed should be well-defined. Failing to do so may lead

the test-taker to focus on another language component such as vocabulary when the intended assessment is of certain grammatical structures.

While retelling a story or news event may appear akin to paraphrasing, the notable distinction lies in the production of longer stretches of discourse belonging to a different genre. Namely, in this type of task, examinees listen to or read a story or news event and subsequently retell it. As with the other tasks, the scoring procedure must align with the desired assessment criteria.

Translation of extended prose is the final example of extensive testing. The translation of words, phrases and brief sentences represents an instance of intensive speaking, while the translation of more extended texts from the native language to the target language exemplifies extensive speaking test formats. The typical examples for longer texts are dialogues and synopsis of a story or a play or a movie. Again, careful consideration should be given to the scoring criteria.

To sum up, a panoply of oral assessment types is available, varying from one word or phrase utterance to a lengthy speech, or from more controlled to the freer tasks. Hence, the assessment of a test type gains importance since the more freedom the test-takers have, the more challenging the scoring is for the assessors.

2.12.2 Scoring procedures for measuring speaking proficiency

For Madsen (1983), facilitating appropriate student responses is only one aspect of the task; scoring the test is equally demanding. Rating scales play a pivotal role in measuring speaking proficiency, providing a structured framework to evaluate various aspects of oral communication. These scales typically encompass a range of criteria that assess different dimensions of language proficiency, including pronunciation, vocabulary usage, grammatical accuracy, fluency, and coherence. Each criterion is assigned a specific rating or score, allowing for a comprehensive and systematic evaluation of an individual's speaking abilities.

The significance of rater training, rubric clarity, and scoring reliability in ensuring valid and fair assessment outcomes have been cited in the literature (Fulcher, 2003; Luoma, 2004; McNamara, 1996). According to Madsen (1983), effective implementation of rating scales necessitates comprehensive training for test administrators and scorers, ensuring that they possess a clear understanding of the criteria and descriptors outlined in

the scale. This training helps mitigate potential biases and inconsistencies in scoring, contributing to the overall reliability and validity of the speaking assessment process. Additionally, regular calibration sessions and monitoring of scoring practices can further enhance the accuracy and consistency of the rating process, ultimately leading to more reliable and informative assessments of speaking proficiency.

One of the notable advantages of rating scales is their flexibility. They can be adapted to suit different language proficiency levels and assessment contexts. For instance, a rating scale for a beginner language test might prioritize basic vocabulary and pronunciation accuracy, while a scale for an advanced language proficiency test might place more emphasis on fluency, complex grammatical structures, and appropriate use of idiomatic expressions. As Madsen (1983) emphasizes that tester can be guided towards more consistent assessments via well-designed rating scales. Similarly, Luoma (2004) underlines the necessity of tailoring rating criteria to specific proficiency levels as it can enhance the validity and relevance of speaking evaluations. Fulcher (2003) also highlights that aligning scale descriptors with task demands can provide correct and equitable scoring outcomes.

The application of rating scales also helps reduce subjectivity in the assessment process (Madsen, 1983). They offer a standardized framework that enables multiple assessors or examiners to provide consistent scores across different test-takers. This objectivity is crucial in ensuring fairness and reliability in language assessment, making it possible to compare the speaking proficiency of various individuals accurately.

One of the notable advantages of rating scales is their flexibility (Madsen, 1983). They can be adapted to suit different language proficiency levels and assessment contexts. For instance, a rating scale for a beginner language test might prioritize basic vocabulary and pronunciation accuracy, while a scale for an advanced language proficiency test might place more emphasis on fluency, complex grammatical structures, and appropriate use of idiomatic expressions.

The application of rating scales also helps reduce subjectivity in the assessment process. They offer a standardized framework that enables multiple assessors or examiners to provide consistent scores across different test-takers. This objectivity is crucial in ensuring fairness and reliability in language assessment, making it possible to compare the speaking proficiency of various individuals accurately.

Moreover, rating scales can be a valuable tool in providing constructive feedback to language learners (Madsen, 1983). By pinpointing specific areas of strength and weakness, learners can better understand how to improve their speaking skills. Educators and test administrators can use the feedback generated from these scales to tailor instructional methods and curricula to address learners' individual needs effectively.

In conclusion, rating scales for measuring speaking proficiency are essential tools in the assessment and development of language skills. They offer a structured and objective approach to evaluating spoken language performance, helping learners, educators, and assessors alike in their pursuit of effective language communication.

In the realm of language assessment, various types of rating scales are used to evaluate oral production, each designed to measure specific aspects of speaking proficiency. These scales help assessors provide a comprehensive and detailed evaluation of a test-taker's language abilities, facilitating a more nuanced understanding of their strengths and areas for improvement. The choice of scale depends on the purpose of the assessment and the level of detail required for scoring and feedback (Weir, 2005). In practice, some common types of rating scales for oral production include holistic scales, analytic scales, and numerical scales (Brown, 2003; Fulcher, 2003; Luoma, 2004; Madsen, 1983).

Holistic rating scales provide a general overview of a test-taker's speaking proficiency, offering a broad assessment of their overall performance without specific breakdowns for different linguistic components. Assessors consider the overall impression of the test-taker's speaking abilities and assign a single score that reflects their overall oral proficiency level. Holistic scales are particularly useful when a quick and general assessment is needed, as they provide a broad evaluation without requiring detailed analysis of individual language components (Madsen, 1983).

Analytic rating scales, on the other hand, break down the assessment into specific components, such as pronunciation, vocabulary, grammar, fluency, and coherence. Each component is evaluated separately, and scores are assigned for each category. Analytic scales offer a more detailed and nuanced assessment of the test-taker's language skills, allowing assessors to identify specific strengths and weaknesses in different aspects of their speaking abilities. This detailed feedback can be valuable for test-takers and educators seeking to target specific areas for improvement (Madsen, 1983).

Numerical rating scales, also known as numerical rubrics, use a numerical system to quantify the test-taker's performance across different language components. These scales typically involve assigning a numerical value to various criteria, such as pronunciation, vocabulary usage, grammar accuracy, fluency, and coherence. By assigning specific scores to each criterion, numerical rating scales provide a standardized and quantitative evaluation of the test-taker's speaking proficiency. This facilitates a more objective and precise assessment, allowing for easier comparison of language proficiency levels across different test-takers (Madsen, 1983).

In sum, the various types of rating scales for oral production serve as essential tools in language assessment, providing assessors with structured frameworks for evaluating speaking proficiency. These scales offer different levels of detail and emphasis on specific language components, catering to diverse assessment needs and helping to provide valuable feedback for language learners and educators.

2.12.3 The assessment of speaking and language teaching methods

The design and use of speaking tests in language education are closely tied to the underlying teaching methodology. Each method reflects and highlights different aspects of language use. Similarly, how speaking is assessed is also influenced by key language teaching methods. To start with, the Direct Method (DM), which emerges in the late 19th century, puts emphasis on spontaneous oral interaction along with no use of first language of the learners. Speaking is regarded as the core skill to be developed and speaking assessment is typically performed in question-and-answer exchanges between the teacher and the student, brief conversations and spoken repetitions. The test types involve oral interviews, dialogs and reading aloud tasks, in which pronunciation, fluency and accuracy are the focus of assessment (Richards & Rogers, 2014).

The second language teaching methodology is Audio-lingual Method (ALM), with roots in behaviorist psychology. The ALM views language learning as habit formation through repetition, drills and pattern practice. Thus, assessment of oral skills involves accurate repetitions of practiced structures instead of spontaneous communication, which puts accuracy, automaticity and structural patterns in the center of assessment. Considering the main focus of the ALM, assessment types typically involve repetition, substitution and transformation drills (Larsen-Freeman & Anderson, 2011).

The Situational Language Teaching (SLT) or also known as the Oral Approach is another language teaching approach which highlights teaching vocabulary and grammar in meaningful situational contexts. The SLT is viewed as akin to the ALM in its structural orientation. Yet, the SLT makes use of more contextualized situations to practice speaking. Therefore, the assessment comprises the correct use of structures in situations through oral drills and structured dialogs (Howatt & Widdowson, 2004).

Fourth approach that uses oral assessment is the Communicative Language Teaching (CLT). The primary focus of CLT is fluency, meaningful interaction, and authentic communication, which regards interactive competence fundamental. Thus, speaking assessments require more performance-based tasks, where fluency, intelligibility and interactional strategies gain prominence. The possible assessment types are problem-solving tasks, role-plays, interviews, and discussions (Canale & Swain, 1980).

Task-based Language Teaching (TBLT) is another major language teaching approach which emphasizes the value of real-world tasks that require students to use the target language for communicative purposes. Thus, regarding the assessment of oral skills, the TBLT pinpoints the successful task competition and communication by focusing on fluency and negotiation of meaning. Regarding this, information-gap and problem-solving tasks as well as simulations are types of assessment in TBLT (Ellis, 2003).

The Post-method Pedagogy, which rejects one-size-fits-all approach to language teaching, is the final approach which moves language teaching from method to post-method. The approach employs speaking assessment in a more context-sensitive and reflective manner, and centers learner-centered and flexible assessment practices by integrating self- and peer-assessment. The focus of assessment is on learner agency, contextual appropriateness and communication strategies; hence, the assessment types include portfolios, presentations, oral journals, integrated performance assessments and self-evaluations (Kumaravadivelu, 2006).

In conclusion, various language teaching methods apply various oral assessment tasks. Depending on the methodological perspectives of the language teaching approaches, these tasks align with instructional goals. To be more specific, while traditional approaches such as the DM and the ALM prioritize accuracy through controlled practice, communicative and post-method approaches emphasize fluency, interaction, and real-life

related tasks. The following section delves into the key challenges associated with assessing the oral skills of second language learners.

2.12.4 The importance of speaking skills and testing speaking

Speaking plays an indispensable role in second language acquisition. It is maybe the most crucial skill for effective communication. It enables interaction and helps to express individuals' feelings and emotions as well as to engage in social interactions. As such, developing speaking skills has become a core component of language education, and the accurate assessment of these skills is necessary to assure language learners' communicative competence.

Perhaps the most significant reason for the importance of speaking skills lies in their central role in communication. Undoubtedly, speaking is the primary means by which individuals convey their thoughts and emotions. It plays a vital role in everyday communication across a range of contexts, including personal, academic, and professional settings. Speaking is regarded as a productive skill since individuals produce the language rather than just receiving it. To be able to produce the language and communicate orally, it is vital to generate verbal output by means of integrating language components like vocabulary, grammar and pronunciation (e.g., Bailey, 2020; Nunan, 1991). It is precisely for this reason that this skill is seen as both important and difficult.

Social interaction is an equally important reason that makes speaking skills crucial. Speaking is a necessity for not only exchanging information but also interpersonal engagement. Thornbury (2005) highlights that in order to participate in conversations, negotiate meaning and build relationships, individuals need social exchange. Via spoken language, individuals can hold dialogues, state their views and respond to others, enhance social connections and mutual understanding. That is, they can engage meaningfully with others. Richards (2008), for example, states that speaking is the primary element for communicative competence, which is essential for successful social exchanges. Nunan (1991), similarly, emphasizes the importance of speaking in communicative language teaching, in which successful interaction, which is necessary for effective communication in social contexts, is highly valued.

Apart from its role in communication and social interaction, speaking skills are fundamental for academic and professional success. Second language learners are often

required to participate in discussions/team works, defend their opinions, and/or deliver presentations in both academic and (future) professional contexts. Thus, effective communication becomes paramount. Swain's (1985) COH highlights the importance of speaking in SLA as one has the opportunity to notice the gaps in their spoken production, which can help learners gain accuracy and fluency in time. Richards (2008) further highlights the importance of speaking fluently and confidently, and emphasizes its impacts on the educational settings and the success in the workplace. As for Ellis (2003), successful oral skills in the professional settings lead to strong interaction, effective decision-making and problem-solving skills. Lightbown and Spada (2021) point out that academic achievement often depends on a student's ability to express his/her viewpoints clearly in presentations, discussions and/or debates. In short, clear and effective communication in educational and professional contexts is often regarded as a key determinant of achievement.

Despite its importance, testing speaking poses numerous challenges due to its subjective and complex nature. One of the main challenges derives from the consistency and reliability in scoring. Louma (2009), for instance, emphasizes that certain factors such as anxiety, the test environment and the interaction between examinees, can cause difficulty in obtaining consistent and reliable assessment results. As a result, some researchers underline the importance of clear criteria and standardized testing procedures (e.g., Hughes, 2020). Fulcher (2003) also states that interactive elements of speaking including fluency and coherence necessitates thorough and well-trained assessors and carefully-designed criteria/rubrics, which are core to valid and reliable assessment.

In addition, types of speaking tests may present challenges. Although different formats of oral tests have been developed so far such as interviews, discussions, role-plays, etc., designing accurate and fair tests for all test-takers may be challenging. According to Brown and Abeywickrama (2010), producing tests that are practical and fair for all examinees as well as reflective of real-life scenarios is difficult. They also note that each test format has its own benefits and drawbacks, so the test items should be in line with the learning objectives and diverse needs of learners.

In a nut shell, speaking plays a pivotal role for successful communication, social interaction and strong academic and professional domains. However, testing these skills poses unique challenges. Therefore, it is vital to design and implement standardized

testing procedures by making use of reliable assessment methods and thorough and standardized test techniques so that language learning can be enhanced on the way to realizing learners' communicative goals.

2.12.5 Major difficulties in testing speaking skills

As highlighted earlier, second language learners often face considerable challenges in developing their speaking skills. Likewise, testing of speaking is also regarded as “the most challenging of all exams to prepare, administer, and score” (Madsen, 1983, p. 147). Thus, some researchers (e.g., Hughes, 2020; Madsen, 1983) seek answers to the possible reasons why testing speaking is deemed challenging. These challenges can be categorized into three main groups: preparation, administration and scoring issues.

To begin with, preparation of the tests can be difficult for test designers. One reason is the interrelation between speaking and listening. Since the speaking requires the interaction of aural comprehension most of the time, the testers should be careful about minimizing the potential impact of aural intake on the learner's performance.

Administration of an oral test is regarded as another difficulty, particularly for examiners. It is almost impossible to implement speaking tests to a large number of examinees due to time issues. It would not be reliable to conduct an oral exam of thousands of test-takers, in which one exam slot lasts about 10-15 minutes.

Thirdly, and perhaps most significantly, is the issue of scoring. Since there are a number of elements of speaking that need consideration such as accuracy, fluency, pronunciation, listening comprehension, the use of correct tone besides grammar and vocabulary, test administrators may find it quite challenging to measure (Madsen, 1983). Furthermore, the intricate nature of the verbal interactions encompasses various dynamics such as turn-taking, initiating, responding, interrupting, topic transitions, etc., making the assessment of speaking proficiency a demanding task to evaluate. Hughes (2020) summarizes oral communication as encompassing both comprehension and active production. Another issue is the elicitation techniques (Brown, 2003). As a tester, when aiming to elicit specific spoken grammatical categories, the stimuli should be crafted to prompt those grammatical categories in a manner that prevents the test-taker from avoiding and paraphrasing, thereby ensuring the production of the targeted form. Finally, challenge in scoring procedures is an issue in testing speaking. Contrary to receptive performance, in

which the answers to the questions are generally predetermined, productive performance tasks necessitate the formulation of question items in a way that aligns with the appropriate scoring and rating procedures. Heaton (2003), for example, states that oral test examiners work under a huge pressure as they need to make fair and subjective judgements in a very short time. Similarly, Harris (1990) believes that “No language skill is so difficult to assess with precision as speaking ability” (p. 81).

In relation to the aforementioned challenge, the training of scorers is also important. Examiners’ familiarity with the rating system is crucial so that they can recognize what is intended in the test task. Heaton (2003) pinpoints that when testers demonstrate consistency and accuracy in scoring, it is advisable for them to conduct real time scoring. Furthermore, test scorers should remain unbiased and not be affected by candidates’ physical appearance or attractiveness.

In sum, many difficulties are at play while testing learners’ oral abilities. These challenges can stem from various reasons. Initially, the preparation of oral tests can be a hurdle for the test designers since it often requires the interaction of aural comprehension. Another challenge is due to administrative hardships, especially with a high number of examinees. Scoring issues can be regarded as a challenge, as well. Since scoring oral skills require many considerations including accuracy, fluency and listening comprehension, testers should approach this issue in a meticulous way. As a result, the challenging aspects of preparation, administration, and scoring often discourage many language teachers and institutions from attempting to assess speaking ability.

2.12.6 Factors that affect speaking performance

Speaking performance of language learners may be affected by various factors. These factors can either be internal or external. Being aware of and understanding these factors is vital for language educators and also researchers so as to form successful teaching and assessment strategies that help different needs of students. These factors can mainly be grouped as linguistic, psychological and environmental factors.

Linguistic factors such as vocabulary, grammar and pronunciation can have a direct impact on learners’ speaking performance. Levelt’s (1989) model of speech production emphasizes the importance of lexical retrieval but remarks it as a vulnerable stage, specifically under pressure. Limited vocabulary knowledge or insufficient grammatical

understanding may cause confusions or misunderstanding while expressing one's thoughts. Goh (2007) indicates that limited vocabulary knowledge notably hinders learners' speaking competence. Ellis (2008) stresses the significance of vocabulary and grammar on forming meaningful utterances, which is key to oral communication. In a similar vein, pronunciation also plays a crucial role in mutual understanding in a conversation. Poor pronunciation skills can lead to misunderstanding, and may affect the speaker's confidence. This type of difficulty may result from the learner's mother tongue (Derwing & Munro, 2005). Whatever the reasons that affect intelligibility and overall speaking performance are, Lewis (2018) notes that integration of pronunciation activities within other language skills can positively impact the learners' speech, and make it more comprehensible.

Psychological factors can impact learners' oral performance. On the one hand, anxiety, lack of confidence, fear of making mistakes and fear of negative evaluations are such factors that often induce communication break-downs. Young (1991) emphasizes the complex and multidimensional aspect of language anxiety as it is shaped by many elements including personal and interpersonal anxieties, learner beliefs about language learning, instructor beliefs, instructor-learner interactions, classroom procedures, and language testing. To this end, Horwitz et al. (1986) emphasize that learners avoid taking speaking opportunities due to anxiety, which, in return, obstructs language development. High levels of anxiety or stress can adversely affect learners' performance and cause to avoid in-class participation (Goh & Burns, 2012). Test anxiety may result in cognitive overload, by extension, reducing students' working memory capacity and impairing their performance (MacIntyre & Gardner, 1994). MacIntyre and Gregersen (2020), on the other side, highlights the role of positive psychology in minimizing language anxiety and promoting oral performance. Positive psychology is the scientific study of human strengths, well-being, and optimal functioning (Seligman & Csikszentmihalyi, 2000), which emphasizes positive emotions such as happiness, gratitude, and hope, positive traits like resilience, optimism, character strengths, and positive institutions such as schools and workplaces. Therefore, to enhance success in speaking, reducing language anxiety via positive psychology gains importance. On the other hand, motivation and willingness to communicate can influence learners' speaking performance. MacIntyre et al. (1998) analyzed the impacts of self-confidence on spoken performance, and learners' willingness

to communicate determined by their self-confidence affect their oral production. Dörnyei and Ryan (2015) talk about the concept of motivational dynamics, a framework for understanding how motivation levels fluctuate and evolve during the language learning process. Namely, it can be assumed that learners with high motivation levels are often more enthusiastic in speaking activities. They are usually open to take risks and practice the target language outside the classroom. Gardner (1985) puts forward the role of intrinsic and extrinsic motivation in second language learning, and emphasizes that the more motivated a learner is the higher levels of proficiency and fluency s/he achieves.

Environmental factors also play a pivotal role in speaking performance. O'Sullivan (2000) remarks that a well-structured, physically comfortable exam environment minimizes external distractions, therefore, help students focus more on their cognitive and linguistic tasks. The classroom environment, for example, can influence oral production. The teacher's supportive approach and the dynamics among the learners can activate active participation and engagement during lessons. Harmer (2001) implies that a classroom where learners who are not afraid of making mistakes and in which mistakes are deemed as a part of learning, facilitates the willingness to communicate in the target language. Furthermore, speaking skills can be developed when learners have meaningful interaction opportunities where authentic communicative situations such as group discussions take place. Likewise, in a testing atmosphere where the learners feel relaxed and less anxious can produce better results in their tests.

In brief, linguistic, psychological and environmental factors can have vital impacts on language learners' speaking performance. Knowing these factors and handling them accordingly can help language teachers, test designers and test administrators create effective language learning and assessment environments., thereby learners are better supported in coping with challenges while speaking.

2.13 Studies on Testing Speaking in EFL Context

The assessment of oral skills in EFL contexts yields some unique challenges as well as opportunities. Effective speaking assessments aid to evaluate learners' true ability in spoken production, yet various factors related to examinees, raters, tasks, criteria/rubric and the physical environment during testing should be taken into consideration, as well. Hence, in order to provide a comprehensible understanding of this phenomenon, previous research on these factors is examined below. When the previous studies are examined, a

bulk of research on examinee-related factors including affective, task-related, personal and linguistic elements are found, and are presented respectively.

In terms of *affective factors*, Woodrow (2006) conducted a research study on understanding second language speaking anxiety, its relationship with language performance, and the primary causes of anxiety experienced by learners. The participants were 275 advanced English for academic purposes (EAP) students enrolled in intensive EAP courses prior to getting in Australian universities. She developed a Second Language Speaking Anxiety Scale (SLSAS) to measure second language learners' anxiety levels, whose validity was confirmed. In addition, speaking anxiety was found to be a significant predictor of oral achievement. Interviews were also conducted to find out the sources of anxiety, with results showing that the most common reason was interacting with native speakers. The finding of the study identified two types of anxious learners, those who experience retrieval interference and those with a skills deficit.

Hanifa (2013) reviewed the literature and compiled the affective factors, particularly the role of foreign language anxiety (FLA) in speaking competency. He tried to address the sources of speaking anxiety, and among them, cognitive, affective and performance-related factors stand out. Cognitive factors include topic familiarity, speaking task genres, the interlocutors and the process itself. Affective factors involve learners' attitudes and emotions towards the topic and/or the interlocutors, including self-consciousness. Finally, performance factors are more concerned about speaking mode, degree of collaboration, the control over discourse, planning and rehearsal time (if there is any), time pressure and environmental factors. Hence, it is crucial to understand the role of these factors on learners' speaking performance and find ways to overcome FLA in EFL settings.

Yılmaz (2015) conducted another study on the impact of test anxiety among Turkish EFL learners. Data gathered from surveys and performance in the speaking tests reflected a strong correlation between test anxiety and reduced speaking proficiency, suggesting that anxiety-reducing interventions could benefit learners to become successful EFL speakers.

Jahbel (2017) examined the factors that affect EFL high school students' speaking performance at Malang, and the problems that they encounter. The sample included 27 participants, comprising 21 students and six teachers, and the data were gathered through questionnaires and interviews. The results gathered from both groups were analyzed separately, and the analysis of the results of the students' data displayed that confidence

was the major factor that affect their performance, which is followed by motivation. Similarly, half of the teacher participants highlighted the harmful effects of anxiety on learners' oral communication skills.

Alrasheedi (2020) also conducted a study on the psychological factors that affect EFL learners' speaking performance with 200 Saudi university students via a questionnaire. The results yield that both internal and external factors impact learners' spoken production. Specifically, these affective factors include shyness, anxiety, fear of making mistakes and peer pressure. Besides, limited vocabulary, lack of exposure to English and insufficient out-of-class speaking practice opportunities are found as other inhibitory factors.

Amoah and Yeboah (2021) investigated the factors influencing the speaking performance of Chinese EFL learners at Nanjing Tech University, their motivational levels and strategies for improving proficiency. Data were collected from 75 participants through a speaking difficulty questionnaire, the English motivational scale and an interview, which was conducted with 10 randomly selected participants. The researchers categorized the factors that inhibit learners' speaking performance into two groups, either linguistic or psychological. The findings displayed psychological factors, including anxiety, fear of mistakes, unwillingness, and fear of negative evaluation, had a more significant impact on speaking performance than linguistic issues like limited vocabulary, pronunciation difficulties, and insufficient grammar knowledge.

A relatively recent study conducted by Jaya, Petrus and Pitolaka (2022) also focused on the relationship between EFL learners' speaking performance, speaking problems and the factors related to these problems at tertiary level. The data, gathered from 67 EFL university level students, were collected through speaking performance test, questionnaire and open-ended questions. The speaking problems addressed by the researchers are affected-related problems such as confidence and anxiety, socially related problems like the scarcity of opportunities to use English outside the classroom and comprehension in speaking lessons, and linguistic factors including vocabulary, grammar, pronunciation and fluency. The researchers also discovered the associated factors for these problems, which are lack of general knowledge, lack of speaking practice, lack of vocabulary and grammar practice, fear of mistake and criticism, low motivation and participation, reading laziness, shyness, nervousness, limited dictionary

use and poor pronunciation skills. The data were analyzed using Chi-Square with $p=603.508$, greater than the critical value (70.003), which yielded a significant association between speaking problems and the factors that lead to these problems.

Dong (2022) explored the factors affecting EFL students' speaking difficulties, specifically regarding linguistic, psychological, and environmental challenges. The research involved 30 EFL learners who answered a questionnaire with 27 items. The findings revealed that psychological factors, particularly low self-confidence, posed the greatest difficulty, followed by linguistic issues such as limited vocabulary for expressing ideas and arguments. Additionally, participants declared a lack of opportunities to practice out-of-class speaking, emphasizing environmental constraints.

When the literature is examined, some studies on *task-related factors* can be found, as well. To start with, Ellis's (2009), also known as the leader of task-based language assessment, longitudinal study revealed that tasks which require higher cognitive engagement led learners to use more complex language structures, and achieve higher success in oral tests.

Lee (2006) analyzed the score dependability of a multitask speaking measure for the new TOEFL test, which consists integrated and independent tasks. Using generalizability theory (G-theory), he examined the data gathered from 261 examinees to determine the impact of task types, raters per task, and subsection lengths on score reliability. Results yielded that increasing the number of tasks, rather than ratings per task, was more effective in enhancing score dependability. Multivariate analyses also revealed high correlations among task-type subsection scores, and the researcher found that increasing listening-speaking tasks slightly improved composite score reliability for fixed section lengths.

Tavakoli (2009) investigated EFL learners' and teachers' perceptions of task difficulty and their criteria for assessing it in oral narrative tasks. 10 EFL learners performed four tasks and were interviewed about perceived difficulty, contributing factors, and how they defined task difficulty. Likewise, 10 EFL/ English for Speakers of Other Languages (ESOL) teachers evaluated the same tasks and discussed factors influencing their learners' challenges. Qualitative analysis revealed that though there were some differences between the two groups' perceptions of task difficulty, there was a significant similarity between them in terms of the criteria they considered in identifying and

defining task difficulty. The findings also support cognitive approaches to task-based language learning and highlight how aspects of Skehan's (1998) model and Robinson's (2001a) triadic framework align with teachers' and learners' perspectives on task difficulty.

Üstün (2010) also investigated the use of communicative language teaching principles and their effects on oral performance. The data gathered through classroom observations and speaking test results showed that integrating authentic tasks improved students' performance and engagement.

Tuan and Mai (2015) investigated the speaking problems of EFL learners at a Vietnamese high school. The data were gathered through questionnaires and classroom observation. The participants included two hundred and three 11th grade EFL learners and ten EFL teachers. The teacher results gathered from teacher questionnaire indicate that students' major problem in speaking is the topical knowledge, which is followed by listening ability, motivation to speak, teacher feedback and confidence, respectively. The results from student questionnaire reflect a similar outcome; that is, topical knowledge is regarded as the most influential factor. Listening ability, confidence and feedback are found to be the other reasons that affect learners' speaking performance. When class observations were analyzed, time for preparation for the speaking task is found to be an important factor in their oral performance.

Like Tuan and Mai (2015), another study conducted by Huang, et. al. (2018) focuses on the impact of topical knowledge on independent and integrated speaking test performance. The design of the study includes independent speaking test performance and integrated speaking test performance to analyze how each was related to topical knowledge. To achieve their purpose, four independent speaking test tasks (independent tasks), and validated four topical knowledge tests (TKTs) were developed and applied by 421 EFL learners. The researchers also invited other 352 students to complete the TKTs, two independent and two integrated tasks. Half of the test takers took the independent tasks and integrated tasks on one topic combination whereas the other half completed tasks on another topic combination. The findings yielded two main results. First, topical knowledge significantly impacted integrated speaking test performance in both groups. Second, the effect of topical knowledge on the two types of speaking test performances (independent tasks and TKTs) was topic-dependent.

Hsieh and Wang (2017) investigated a range of fluency, grammar, vocabulary, and content features of young EFL learners' speaking performances, using a discourse-analytic approach. Responses to the speaking section in the TOEFL Junior® Comprehensive test gathered from 179 test takers were analyzed via Mixed-design ANOVAs. The researchers compared the oral performance across four proficiency levels and two task types, a picture narration task and an integrated listen/speak task. It was concluded that the discourse features largely determined the proficiency levels of test takers. In addition, the findings presented task types' impacts on measures of grammar, vocabulary, and content, but no effect on features of fluency.

O'Grady (2019) explored how varying lengths of pre-task planning time affect performance in speaking tests for university admission. The participants were comprised of 47 Turkish-speaking English learners, who were divided into proficiency groups based on a placement test, completed four monologue tasks: two picture-based narratives and two descriptions. Various planning times including 30 seconds, 1 minute, 5 minutes, and 10 minutes were assigned. Scores, awarded using an analytic and a binary-choice rating scale, were analyzed with multifaceted Rasch measurement. The results revealed that planning time effects varied by rating scale, task type, planning duration, and proficiency level. Also, increases in scores were observed more on the picture-based narrative tasks than on the two description tasks. The 5-minute planning condition yielded the largest overall score improvements.

Another study by Crossley and Kim (2019) examined the effects of text-based relational (e.g., cohesion), propositional (e.g., lexical), and syntactic features in a source text on the subsequent integration of the source text in oral responses. The effects of vocabulary integration on people's evaluations of speaking performance, taking into account test-takers' individual characteristics (e.g., listening proficiency, age, GPA, working memory capacity) and test-taker strategy use (e.g., note-taking strategies) were also investigated. Speaking samples from a total of 263 test takers were collected using The Internet-based Test of English as a Foreign Language (TOEFL-iBT) survey forms with integrated listening/speaking items. Linguistic features, together with individual features and note-taking data, were used to predict participants' speaking proficiency scores. The results revealed that the linguistic features of the source text are an almost perfect predictor of which words test-takers will integrate into their oral responses. Furthermore, text

integration was found to be a significant factor which influence participants' speaking proficiency scores, beyond test takers' individual characteristics and note-taking strategies.

Rater-based studies can be summarized as follow. To begin with, Knoch (2011) conducted an empirical study on rater training programs. In her study, the control group received an intensive training on a speaking test rubric, whereas the experimental group did not. She found that rubrics with specific descriptors lessened the rater subjectivity. Also, the results highlighted that the more the structured training sessions is, the more the rater reliability is.

Winke et. al. (2012) investigated whether familiarity of test-taker's accent leads to rater bias. The participants included two groups: first group consisted 107 examiners who are L2 speakers of Spanish, Chinese, or Korean, and the second one consisted of test takers who were native speakers of Spanish, Chinese and Korean, 24 people each, for a total of 72 test takers. The raters examined 432 TOEFL IBT™ speech samples, and the researchers analyzed the ratings via a multifaceted Rasch measurement approach. The findings yielded that both L2 Spanish and L2 Chinese raters were significantly more lenient with L1 Spanish and L1 Chinese test takers, respectively. The researchers implied that rater training should address raters' linguistic background as a potential rater effect.

Eckes' (2015) study explored rater effects and biases in assessing German EFL learners' performance. By analyzing rater scoring patterns and conducting interviews, Eckes highlighted how individual raters' characteristics influenced speaking test outcomes and suggested methods for mitigating these biases.

In his dissertation, Han (2020) searched for the effects of rater characteristics, specifically rater expertise, working memory capacity (WMC) and cognitive functionality on scoring second language speaking performance. The participants were from the US and the UK, totaling 90 raters, who took part in the study in two phases. The aim of the first phase was to determine the contributions of rater expertise and WMC to their scoring performance, while second phase mainly focused on determining the strategies raters made use of during assessment in cognitive lab sessions. The results of the first phase yielded a significant effect of rater expertise on scoring accuracy. That is, the more experienced the raters, the more accurate the rating scores were gathered in the speaking tests.

In Sourı and Merç's (2020) study, although the main focus was on investigating the effects of self-correction of errors during the TOEFL-IBT speaking test through awareness raising, the participant in the study also emphasized the importance of rater feedback. Specifically, the feedback given by the raters after each test implementation contributed to the participant's subsequent speaking performances, for it offered an opportunity for the learner to identify mistakes and self-correct the errors. The findings indicated an overall improvement in speaking skills, especially regarding vocabulary range, elaboration, and using discourse linkers, as noted by the raters.

Another study by Neittaanmäki and Lamprianou (2024) examined the changes in rater severity and consistency, focusing on the L2 Finnish-speaking subtest of Finland's National Certificates of Language Proficiency between 2009 and 2019. Using data from 45 rating sessions involving 104 raters and 59,899 examinees, the study analyzed trends via the Many-Facets Rasch model and generalized linear mixed models. Findings of the study demonstrate that raters became slightly more lenient over time. In addition, factors such as the composition of the rater community, lead examiners, rating and training methods (on-site or remote), and changes in rater practices significantly influenced rating behavior, highlighting the impact of systemic changes on assessment reliability.

Some studies focus basically on the *criteria and rubric* use. Karslı (2002), for example, conducted research at Anadolu University SFL with 5 EFL teachers on the use of two different analytical criteria while assessing 36 students' oral performances. The results of the study indicated a positive attitude towards four-band criteria rather than five-band one since the participants found the descriptors of five-band criteria vague. Also, the interview results yielded some common problems about the four-band criteria, which are mainly about the descriptors not being clear enough, students' performance' incompatibility with the criteria, the problems of the bands, particularly with the highest and the lowest ones, and the big value differences among the bands. The researcher also recommends test evaluators to take training and attend norming sessions before speaking test implementations.

Fulcher and Davidson (2016) conducted another empirical study in which they analyzed the effectiveness of various rating scales used in EFL contexts. They found that rubrics that have specific descriptions lessened the rater subjectivity.

Another empirical study by Khabbazzbashi and Galaczi (2020) examined the impact of holistic, analytic, and part marking models on scoring a semi-direct online speaking test. Results showed that scoring models significantly influenced Common European Framework of Reference (CEFR) level assignments, with holistic marking assigning higher levels and part marking enabling finer distinctions. Part marking demonstrated superior measurement qualities, highlighting its potential for improving the validity and fairness of speaking test scores.

Apart from these factors, some studies related to the effect of *computational* scoring compared to human scoring were conducted. Zechner and Evanini's (2016) empirical study, for example, found that automated systems could reliably score certain aspects of oral language like pronunciation and fluency. However, they also underlined some limitations in assessing more subjective elements such as conversational engagement. In similar study conducted by Burstein et al. (2013), the efficacy of automated scoring in large-scale assessments was tested, and the results yielded high correlations with human ratings.

Alzahrani (2020) compared two modes of oral proficiency testing, which are Oral Proficiency Interview (OPI) and Video Oral Communication Instrument (VOCI) with regard to test-takers' complexity, accuracy and fluency. OPI is a direct-testing mode whereas VOCI is indirect. The accuracy and the fluency level were found higher in OPI. It also explored the participants' perceptions on these two modes, and the results reflected a favor on OPI among participants since they felt more comfortable with this testing mode.

Besides, examinees perform a speaking task under various conditions. Thus, *context-based* conditions, such as the peer effect, physical conditions of the room, duration of the test and the condition of being-video recorded can play a role on the scores of the test-takers and the test-givers, as well. In her master's thesis, Kwon (2015) searched the effects of paired speaking on Korean middle school EFL learners via mixed methods research including questionnaires, retrospective protocol sessions and interviews. The findings mainly showed a positive perception towards the paired format of the speaking exam since they could speak with confidence. Nonetheless, some students deemed pair format tests as pressure and anxiety evoking as they felt it decreased the motivation and increased the competitiveness. Beside peer effect, another context-based condition, which is related

to performance conditions, were revealed in Tuan and Mai's (2015) study. The findings of this study disclosed the effect of performance conditions including time pressure as one of the factors which impacts students' speaking performance. Namely, the study asserts that time limitation might lead the students to do less than optimally.

In addition, there are some studies whose focus is mainly on the *readiness* for the tests. Duran (2011), for example, investigated the EFL teachers' and learners' attitudes towards the washback effects of classroom-based speaking tests. At the time of data collection, all participants, 307 preparatory class students and 45 instructors, were either studying or working at a state university in Türkiye. Though the main purpose of the study was to reveal the effects of tests on teachers' and students' classroom applications, one of the findings comes to the fore. That is, participants verbalized the significance of preparedness for speaking tests since it enhances their general speaking skills, as well.

Moreover, Qasemi (2020) examined the factors that affect students' speaking performance in an English department with 108 language undergraduate learners. The results showed that many students perceive learning English as important and necessary, but have some speaking problems as they do not have the opportunity to practice outside the class, which influences their oral communication ability in English.

Some studies reveal the effects of EFL learners' *linguistic competencies*, generally involving vocabulary, grammar, and pronunciation. For example, Afebri and Muhsin (2019) explored the factors affecting EFL learners' oral performance. They categorize these factors as performance conditions (time pressure, planning, standard of performance and amount of support), affective factors (such as motivation, confidence and anxiety), listening ability, topical knowledge and feedback during speaking activities. Their aim was to analyze these factors and determine the difficulties in speaking performance of the 10th-grade students in Indonesia. The researchers employed a quantitative design in the form of survey conducted by 30 students, and analyzed by a statistical package. The results of this research revealed that among the factors linguistic factors including grammar, vocabulary and pronunciation were ranked the highest.

Al Hosni (2014) also investigated speaking difficulties encountered by fifth grade EFL learners. The analysis of his study displayed that the major challenges in oral production were related to linguistic difficulties, mother tongue use, and inhibition. He concluded that students were unable to speak in English due to lack of vocabulary and grammar

structures as well as lack of sentence formation skills derived from their mother tongue use. In particular, the results yielded that majority of the learners (86.7%) found it difficult to form complete sentence during speaking. Also, 76.7% of the students agreed that they had difficulty in pronouncing some words accurately, and 73.4% explained having difficulty in speaking English because of limited understanding in grammar. In addition, 70% found it challenging owing to limited vocabulary.

Some of the studies also focus on more than one factor. To illustrate, In'nami and Kouzami (2016) conducted a study to find out the percentage of variation in both L2 speaking and writing performance by focusing on the tasks, raters and their interaction. For L2 speaking, they utilized 28 datasets from 21 studies, and the findings reflect that most of the score variances are defined by examinees' performance. Also, they deduced that task and task-related interaction effects explained more of the score variances than did the rater and rater-related interaction effects. However, they also concluded that context, scoring methods and criteria may vary performance on the tasks.

Jahbel (2017) investigated the factors that affect EFL high school students' speaking performance with 21 students and six teachers. The data were gathered through questionnaires and interviews. The analysis of the results displayed various factors that affect their performance, yet all of the teacher participants agreed on the "time for preparation" and "motivation" as the most impactful element on students' speaking performance. 66% of the teachers indicated the negative effects of topical knowledge and confidence as influencers on oral performance. Furthermore, time allowed to perform a speaking task, pressure to perform well and anxiety were also acknowledged by half of the participants. Listening ability and the listeners' support were deemed as the least effective factors, though. The majority of the students (71%), on the other hand, found "confidence" as the biggest impact, which was followed by motivation (57%), time for preparation (57%) and topical knowledge (57%) as the second most influencing factors. 40% of students considered listening ability and listeners' support as a factor. On the contrary, just a small number of the students 33% considered anxiety and pressure to perform well considered as an affecting factor on their performance.

Moreover, Jaya, et.al. (2022) conducted research on the effects of EFL university students' speaking performance, speaking problems, and factors associated with these problems. More than half of the participants did not perform well in the speaking test.

The speaking problems were grouped as affected-related problems including self-confidence and anxiety, socially related problems like difficulties to find opportunities to learn English and comprehension in speaking class, and linguistically related problems such as fluency, grammar, vocabulary, and pronunciation. The data were calculated via Chi-Square, and the results yield a significant association between the speaking problems and the factors causing the problems, which are general knowledge, lack of speaking practice, fear of mistake, lack of words usage and grammar practice, low motivation, low participation, reading laziness, shyness, less dictionary usage, nervousness, fear of criticism, and unfamiliar words pronunciation were found to be the basic reasons that lead to speaking problems.

Another study conducted by Tong (2022) examined the factors influencing the speaking skills of EFL learners at the University of Phan Thiet and identified key elements of effective English communication. Data gathered from 175 participants via a questionnaire revealed two primary categories of challenges, which are internal factors, including low self-confidence, listening ability, and lack of ideas, and external factors, such as performance conditions, teaching methods, and learning environment. The researcher also emphasized the importance of diverse vocabulary, good pronunciation, and fluency for good communication.

3 METHODOLOGY

3.1 Research Design

This study employed a mixed-method design, specifically convergent mixed method design, in which both quantitative and qualitative data were collected and analyzed simultaneously (Dörnyei, 2007). The results were merged and compared, and the discrepancies were explained by evaluating whether the results corresponded or showed divergence. In this design, equal priority is given to both types of data, allowing for a comprehensive interpretation of the mixed results obtained from both methodologies so that one type of data collection method can compensate the weaknesses of the other (Şahin & Öztürk, 2019).

3.2 Research Setting

Since the aim of the study is to find out the factors that influence English as a Foreign Language (EFL) learners' oral performance, Anadolu University School of Foreign Languages (AUSFL) was regarded as one of the most convenient settings for the data collection. AUSFL is particularly suitable for several reasons. First and foremost, it provides a representative environment in which the research topic can be explored. The institution offers a diverse student population and a curriculum aligned with national educational standards, which could enhance the generalizability and relevance of the findings. In addition, the researcher was employed at the institution at the time of data collection, which presents several practical and methodological advantages. Being an insider allows the researcher to access the participants, with whom she has established rapport with, more easily, which was especially valuable for gathering in-depth insights. Furthermore, the researcher's familiarity with the institutional context helps to anticipate and manage potential challenges, thus contributing to the overall feasibility and effectiveness of the study. Ethical considerations were carefully taken into account to avoid bias and maintain objectivity in data interpretation. Thus, AUSFL provided an optimal setting for conducting the present study, offering both practical accessibility and contextual relevance.

In AUSFL, students are placed into five proficiency levels, which are categorized as D, C, B, A, and A+, respectively. Level D is equivalent to beginner level, whereas Level A+ corresponds to upper-intermediate level. This structured progression allows for a clear

and systematic approach to language instruction. Learners at all levels are required to take a midterm and a final exam each semester, which consists of three main components: a booklet exam in which listening and reading skills as well as language use (grammar, vocabulary and function) are assessed, a writing exam, and a speaking exam. The Booklet exam consists of three sections: Listening, Reading, and Grammar & Vocabulary, and includes multiple-choice, matching, multiple matching, and error-identification question types. It is administered in a single session at the announced date and time. Each section includes questions aligned with the topics covered in class and the same learning outcomes. The exam begins with the Listening section. After the Listening section, students are given a certain amount of time to complete the remaining two sections. Students who arrive late may join the exam within the first 15 minutes after the Listening section ends. In this case, their exam will be evaluated based only on the remaining two sections. The Writing exam is administered in a separate session at the announced location, date, and time. Students are given a written production question prepared in line with the learning outcomes of their level, and they are expected to follow the instructions related to the question carefully. In the designated section of the exam paper, students may take notes and prepare before starting to write. This section is not included in the grading, though. A suggested word limit is provided on the exam paper for students to follow while answering the questions. Staying within this range can help prevent insufficient or overly detailed/irrelevant explanations. Papers are graded out of 20 by two different instructors independently using an analytical assessment rubric (the AUSFL Writing Criteria, which is also used in the evaluation of assignments, is applied in this process). The Speaking exam is conducted by a panel of two instructors. Each interview is allocated 15 minutes. Students take the exam in pairs, and the entire exam process is recorded on video. The exam venue and time are announced to students in advance through lists. The speaking exam questions are prepared in accordance with the speaking learning outcomes of the relevant level. Students' speaking performances are assessed independently by the instructors on the panel after the students leave the exam room, using an analytical assessment rubric (AUSFL Speaking Criteria is used in this process). The evaluation is done out of 20. The average of the two instructors' scores constitutes the final speaking exam grade. This score, expressed out of 20, is then converted to a score out of 100 and entered into the system. Except for level D, which is the lowest language level, the Speaking exam consists of two parts and a total of three questions.

Part A asks individual questions, in which each student is asked two questions and is expected to respond individually. No additional time is given to think before answering. Students are expected to speak for 1-2 minutes per question. Part B has a discussion question. In this part, students are given a discussion topic. After one minute of preparation time, they are expected to engage in a 3-5 minute discussion. Students are NOT expected to reach a consensus on a single option at the end of the discussion. However, it is important that they provide a well-rounded evaluation of the topic appropriate to their level. In Level D, there are two parts, yet with different aims. Part A requires spoken production, in which each student is asked two individual questions, and they are expected to respond individually; and Part B requires spoken interaction, where students are given a common question and are expected to interact with each other and exchange ideas related to the question. With this assessment structure, learners' language development throughout the program can be observed consistently.

3.3 Participants

Since the participants were convenient to the researcher and were available for the study (Creswell & Guetterman, 2019), they were chosen through convenience sampling, and consist of mainly two parties: language instructors and EFL learners. As shown in Table 3.1, since the participants were determined on voluntary basis, the number of the participants displayed variety regarding the research instruments.

Table 3.1

Number of Participants during Data Collection

		Quantitative	Qualitative	
		Questionnaire	Interview	Stimulated recall
Students		210	15 (3*5 prof. levels)	15 (3*5 prof. levels)
	Regular	86	15 (3*5 prof. levels)	9
Instructors	Test designers	10	5 (1*5 prof. levels)	-
	Administrators	2	2	-

First of all, the quantitative data were gathered via a questionnaire, which was given to all stakeholders at AUSFL including students, regular instructors, test designers and administrators. The questionnaire was responded by 210 EFL learners, and 98 instructors. Among these instructors, 86 were regular instructors actively teaching at the time of data collection, 10 were test designers, and two were deputy directors at AUSFL. These 10 test designers were determined to represent all language proficiency levels at the institution. Namely, two individual test designers were selected from each of the five different levels, ensuring balanced representation across the full range of language levels. Table 3.2. represents the demographic information of the instructors.

Table 3.2

Instructors' Demographic Information (Questionnaire)

	No. of participants (N=98)				
Gender	female	male	prefer not to say		
	67	30	1		
Age	51-60 years old	41-50 years old	31-40 years old		
	12	51	35		
Educational background	PhD	MA	BA		
	6	49	43		
Teaching experience	over 15 years	11-15 years	6-10 years		
	70	20	8		
Teaching level	A+	A	B	C	D
	16	23	17	32	10

As can be seen in the table above, the majority of participants (n=67) were female, along with 30 males. In terms of age, almost half of the participants fell within the age range of 41 to 50. Additionally, 35 participants were aged between 31 and 40, while the remaining participants were distributed between the ages of 51 and 60. An analysis of the participants' educational background reveals that 50% of the respondents possess a Master's degree (n=49), while 43 are undergraduates. The remaining participants hold a PhD degree (n=6). Besides, a significant majority of teachers have more than 15 years of teaching experience. Specifically, 70 participants have been teaching over 15 years, 20 have been teaching for 11 to 15 years, and a smaller fraction, 8 participants, possess 6 to 10 years of teaching experience. Regarding their teaching level at the time of data collection, out of the 98 teachers surveyed, 32 participants were found to be teaching at level C comprising the majority, followed by 23 instructors at level A, 17 at level B, 16 at level A+, and 10 at level D, arranged in descending order from the highest to the lowest.

Table 3.3 depicts the demographic information including gender, age and English learning experience along with the current language proficiency levels of the student participants. Since the data were gathered in the Spring semester, some of the students who successfully completed Level A in the Fall semester had the opportunity to start in their departments. Thus, the number of the students in the second semester decreased. When the 2023-2024 Spring Term Speaking midterm exam was examined, only 442 participants took the exam, and almost half of these students completed the questionnaire.

Table 3.3

Students' Demographic Information (Questionnaire)

	No. of participants (n=210)				
Gender	<u>female</u>	<u>male</u>	<u>prefer not to say</u>		
	136	72	2		
Age	<u>16-20 years</u>	<u>21-25 years</u>	<u>more than 25</u>		
	175	28	7		
English learning experience	<u>over 15 years</u>	<u>11-15 years</u>	<u>6-10 years</u>	<u>2-5 years</u>	<u>Less than a year</u>
	5	31	53	38	83
Current level	<u>A+</u>	<u>A</u>	<u>B</u>	<u>C</u>	<u>D</u>
	26	50	83	35	16

As indicated in Table 3.3, similar to the instructors' numbers, the majority of the participants (n=136) were female, with 72 males. 2 participants did not want to state their gender. Also, most of the participants fell within the age range of 16 to 20. To be more specific, 175 participants were between 16-20 years old. Additionally, 28 participants were aged between 21 and 25, while the remaining participants (n=7) were distributed between the ages of 26 and above. Moreover, an analysis of the participants' educational background reveals that majority of the respondents (n=83) have been learning English for a year. 53 students have been learning the language for 6 to 10 years whereas 38 learners have been engaged in the language for 2 to 5 years. Only 31 students have been learning the language for 11-15 years, and the remaining participants (n=5) have been learning English for more than 15 years. When the number of the students were examined closely, there was a clustering at levels A and B, 192 of these learners were studying at Level B and 137 were at level A. Level C comprised of 90 students which is followed by Level A+ with a number of 81 students. Level D is the repeating level, thus has the lowest

number of students, namely 32. To be more specific, among 210 students, 83 were studying at level B, 50 at level A, 35 at level C, 26 at level A+ and 16 at level D.

For the qualitative part of the study, 15 instructors were selected through random sampling, specifically three from five language levels who graded students' oral performances in the achievement tests. These instructors were interviewed in the Fall term of 2023-2024 academic year. Besides, one Testing Unit member from each level, totaled five, and two administrators were interviewed. Namely, in terms of interviews with the teachers, 22 participants were volunteered to take part. Plus, in the Spring term of the same academic year, one week after the implementation of the speaking exam, the video stimulated recall protocols (VSRPs) were conducted with nine instructors, who were not only actively teaching but also administering the speaking exams. Except one language level, two instructors from each level were volunteered to take part in VSR interviews. These instructors, also the administrators of the exam, watched 15 students' speaking exam videos (See Table 3.1).

As in instructor interviews, three students were chosen from each of the five different levels; thus, the interviews were also conducted with 15 students across five language levels after they were informed about their speaking exam scores. Additionally, VSRPs were also conducted with these 15 students so that a balanced representation across all of language levels could be ensured (See Table 3.1).

3.4 Data Collection

This current study is represented in a mixed methods research (MMR) design, in which both quantitative and qualitative data were gathered simultaneously. Creswell (2012) puts an emphasis on the importance of the combination of both quantitative and qualitative research methods to better understand the phenomenon. The present study employs a convergent parallel mixed method (MM) design, concurrently collecting both quantitative and qualitative data during the same data collection period (Dörnyei, 2007). In this design, equal priority is given to both types of data, allowing for a comprehensive interpretation of the mixed results obtained from both methodologies so that one type of data collection method can compensate the weaknesses of the other (Şahin & Öztürk, 2019). In the following part, the main points in MMR design will be explained.

3.4.1 Mixed methods research design

Mixed methods research (MMR) is a research design which incorporates both quantitative and qualitative research techniques in one study. Heighem and Crocker (2009) describe the term ‘mixing’ as “the data or the findings are integrated and/or connected at one or several points within the study” (p. 137). Similarly, several definitions of MMR exist in the literature. Initially, as cited in Creswell and Guetterman (2019), “A mixed methods research design is a procedure for collecting, analyzing, and “mixing” both quantitative and qualitative methods in a single study or a series of studies to understand a research problem”. In a similar way, Saldaña (2011) defines MMR as the integration of qualitative and quantitative methods in one study, in which both forms of data complement each other to address research questions. Tashakkori and Teddlie (2003) also point that MMR is not only a mixing of two techniques but a philosophical stance that involves the deliberate blend of quantitative and qualitative approaches to enrich research findings. Brown (2015) further draws attention to what MMR is not. Namely, he advises researchers to consider what MMR is not by stating “MMR is not just a hodge-podge of qualitative and quantitative methods combined with no rhyme or reason” (Brown, 2015, p. 79). Namely, he emphasizes the systematic and complementary manner of blending both qualitative and quantitative perspectives in a research project (Brown, 2015). Yet, all of these definitions underscore the reasons of combining both qualitative and quantitative methods. Collecting data from different approaches and merging different data sources reinforce each other; therefore, interpreting the merged data can provide a comprehensible understanding of a research problem.

In MMR, both numeric information such as data gathered via questionnaires and narrative information such as data collected from interviews, picture descriptions, and so on are collected to provide more insightful responses to the research inquiries. As Saldaña (2011) states, the combination of qualitative and quantitative data helps researchers delve deeper into the data compared to using either method alone. Apart from providing a better understanding of the research problem than either type by itself, Creswell and Guetterman (2019) suggest other reasons for MMR design. For example, MMR study can be utilized when one approach is not enough to answer the research problem, or “to provide alternative perspectives in a study” (p. 545). Likewise, according to Johnson and Onwuegbuzie (2004), the reasons to use mixed methods (MM) is to explore different

dimensions of a problem, triangulate findings and provide a comprehensive explanation. Likewise, according to Tashakkori and Teddlie (2003), MMR is ideal for the researchers who would like to address complex research questions and seek validation of findings as well as researchers who are interested in interdisciplinary studies in which the merge of qualitative insights and quantitative components are needed.

Beside the reasons for the necessity of MMR, the types of MM designs deserve to be probed. Johnson and Onwuegbuzie (2004) advise researchers of MMR design to decide on the dominance and the time of implementation, displayed in Figure 3.1. Dominance refers to whether quantitative or qualitative data are dominating, whereas time of implementation deals with whether qualitative and quantitative data will be implemented sequentially or concurrently, and are displayed with symbols. The symbol “+” refers to the simultaneous implementation of both qualitative and quantitative studies, whereas the symbol “→” refers to two data sources’ being conducted in a sequential order. Also, the capital letters are used to imply the priority and weight in the research, while the lower case means the opposite. Nine variations of mixed methods regarding the time of implementation and dominance are seen in the figure below.

Figure 3.1

Mixed-methods design matrix with mixed-method research designs shown in the four cells

		Time Order Decision	
		Concurrent	Sequential
Paradigm Emphasis Decision	Equal Status	QUAL + QUAN I	QUAL → QUAN QUAN → QUAL II
	Dominant Status	QUAL + quan QUAN + qual III	QUAL → quan qual → QUAN QUAN → qual quan → QUAL IV

Reference: Johnson & Onwuegbuzie, 2004, p. 22.

Creswell (2012) also puts emphasis on the classification of MM studies, and suggest researchers to answer some questions before selecting the type of MMR, which are related

to intent, timing and emphasis. The first question asks the intent for combining the quantitative and qualitative data. The researchers need to determine their intentions and decide whether the idea is to compare the two databases, to validate one database, or to gain a more complete understanding of a problem (Creswell & Guetterman, 2019). The second question delves into the timing, specifically the sequence of collecting the quantitative and qualitative data. The third one probes the emphasis regarding priority or weight to the quantitative and qualitative data collection, though equal emphasis is sometimes possible. Despite several models, (Creswell & Plano Clark, 2018; Tashakkori & Teddlie, 1998), by using these three questions, Creswell and Guetterman (2019) underline three most frequently used MMR models in educational settings: the convergent design, explanatory sequential design and exploratory sequential design (Figure 3.2). Following is a brief description of the main characteristics these designs.

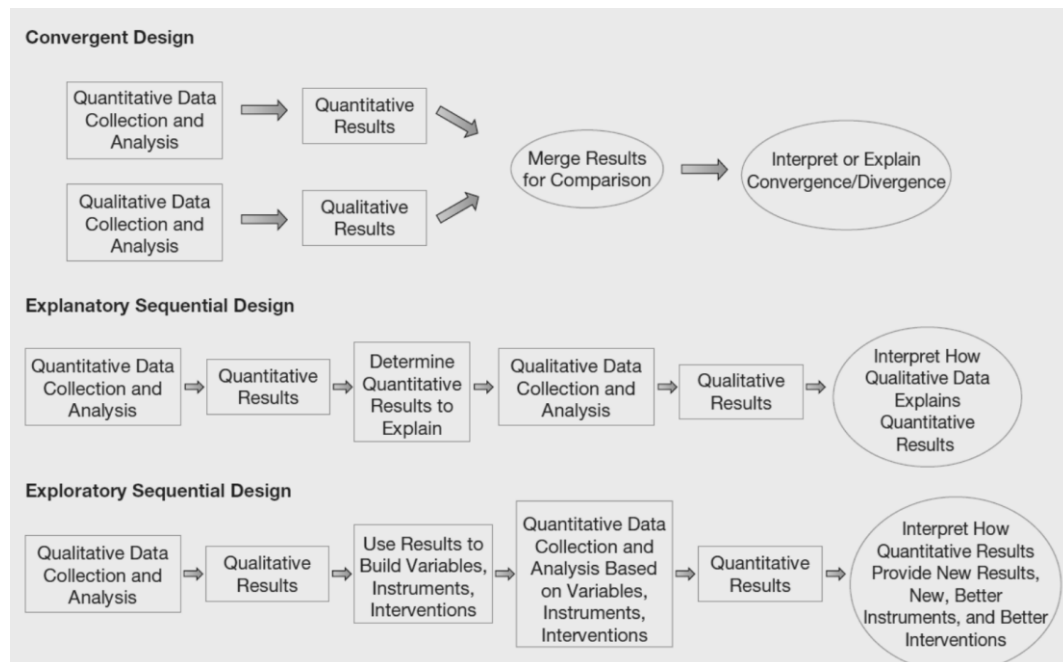
In the convergent design, also called parallel design, both qualitative and quantitative data are gathered together simultaneously, the results are merged and compared, and discrepancies are explained, if any. The main rationale for this design to check one another, and compensate each other's limitations. That is, the researcher collects both qualitative and quantitative data at the same time, conducts separate analyses for each data set, and makes an interpretation by evaluating whether the results from both data analyses correspond or show divergence.

In an explanatory sequential MM design (or a two-phase model), first quantitative data and then the qualitative data are gathered in order to better explain the quantitative results. This design is preferred when the quantitative data and results provide a general picture of the research problem, and more elaborate analysis is required to be able to explain the quantitative data set with the help of qualitative data. As suggested by Johnson and Onwuegbuzie (2004), the researcher prioritize *QUAN*, which is followed by a small *qual* element (Creswell & Guetterman, 2019).

Contrary to explanatory sequential MM design, in an exploratory sequential MM design, first the qualitative data and then the quantitative data are collected. A basic rationale for this design is to explain the relationships found in the qualitative data with the help of quantitative data. A typical way of applying this design starts with examining a phenomenon, determining the themes, and finally designing an instrument and testing it. Mainly the emphasis is put on *QUAL* more than *quan* (Creswell & Guetterman, 2019).

Figure 3.2

Basic mixed methods design



Reference: Creswell & Guetterman, 2019, p. 552.

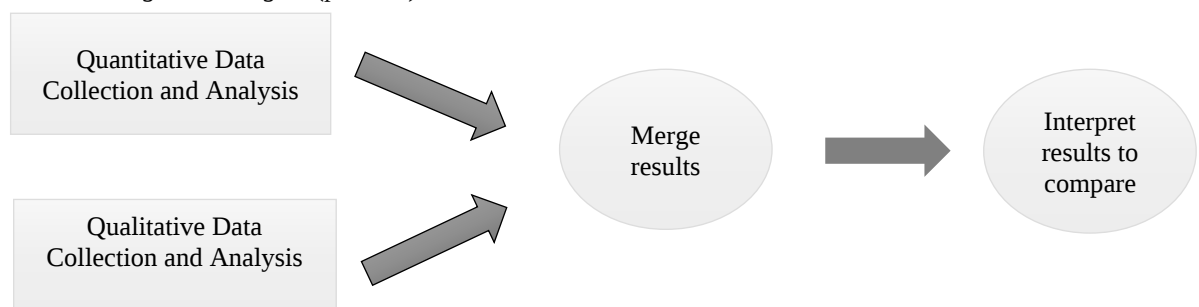
Apart from basic MM designs, there are other MM designs that are often used in the educational field, which are experimental, social justice and multistage designs (Creswell & Guetterman, 2019). To start with, the experimental design, as the name suggests, adds an experiment to the qualitative data before, during, or after the experiment. This design allows to collect data either concurrently or sequentially, which makes the basic design either convergent, explanatory or exploratory. Additionally, Creswell & Guetterman (2019) suggest a relatively new MM design called social justice design. As in the experimental design, in social justice design, either a convergent, explanatory or exploratory way of analysis is utilized as well to address an issue related to social justice. Thus, a transformative framework is sought with an emphasis on a social issue related to feminism, race, ethnicity and so on. Multistage design is another complex design which grounds on either convergent, explanatory or exploratory sequential designs. This design is utilized when the impact of a program or a project is intended to be sought. Therefore, it encompasses not only formative but also summative assessments, and builds on carefully designed stages. The staging first starts with needs analysis, and followed by the development of a theory, instrument design and testing it, respectively.

Although different MMR designs are cited in the literature, the invaluable framework for addressing more complex research questions provided by MMR designs cannot be overlooked. Triangulation of the data results and better understanding of the phenomenon is possible with the integration of both qualitative and quantitative data, which provides breadth and depth to the study. As noted by many scholars including Saldaña (2011), Creswell (2012, 2014), Brown (2015), and Creswell and Guetterman (2019), MMR design is a powerful tool in enriching the findings and helping researchers gain better insights of a research problem.

In this study, Creswell's MM design, particularly convergent parallel mixed methods design was preferred due to its structured yet flexible approach, which is suitable for investigating the complex nature of the research question (See Figure 3.3). Creswell's (2014) emphasis on the simultaneous collection of both quantitative and qualitative data in a convergent design aligns with the purpose of this research to capture deeper contextual insights. This design also allows for the validation of findings through triangulation since it integrates statistical analysis and narrative exploration, which strengthens the reliability and depth of the results. Consequently, providing a holistic view of the phenomenon makes this design a suitable choice for this study.

Figure 3.3

Research design: Convergent (parallel) mixed methods



Reference: Creswell & Guetterman (2019, p. 552).

3.5 Data Collection Tools

Keeping the implementation in mind, the quantitative data were collected via a questionnaire developed for both instructors and students to find out the major factors that might have an impact on learners' oral performances in the achievement tests.

Although the items in the questionnaire given to all stakeholders were the same, the items in the demographic information part differed.

The qualitative data, however, were collected via two different tools. First of all, interviews with the instructors, test-designers, administrators and the students were held. Also, stimulated recall protocols, which were implemented with the instructors who are at the same time the raters of the speaking exam one week after the implementation of the exam. The rater instructors watched the exam recordings and were asked to specify the factors that influence their grading of the learners' oral performance.

3.5.1 Questionnaires

In order to find out the factors that affect EFL learners' speaking scores through the eyes of instructors and students, a questionnaire named *Factors Affecting Students' Speaking Performance (FASSP)* was developed by the researcher (See Appendix A for instructors & Appendix B for students). The purpose of this questionnaire is to determine the major factors that have an impact on students' oral performance. To achieve this purpose, a 4-point Likert scale was developed, with an open-ended question which probes into other possible factors. The FASSP was designed to collect structured and standardized responses, which enabled the measurement of participants' attitudes, perceptions, and levels of agreement regarding the focus of the study (Joshi et al., 2015). The scale was designed to avoid a neutral midpoint, allowing participants to take a clear position. Additionally, two versions of the FASSP were developed: instructor version and student version. While Part II, which examines factors affecting students' speaking performance in achievement tests, remained identical, Part I, designed to collect demographic data, differed slightly. In addition, to ensure clarity and comprehension, the students were given the Turkish version of the questionnaire. The responses to the both versions of the questionnaire were obtained through Google Forms, which is an online survey administration software.

Initially, the first draft of the questionnaire items was examined together with two experienced EFL teachers. Later, expert opinions were gathered from eight experts, five of whom are professors at Anadolu University English Language Teaching (ELT) department and three of whom are experienced EFL teachers and testers at AUSFL. The feedback from the experts were obtained via Google Forms in which all items were written one by one which is followed by three questions: appropriate for use, not

appropriate for use and appropriate for use after necessary changes. Also, a comment box was provided after each item so that the experts could state their opinions when they checked *not appropriate for use* or *appropriate for use after necessary changes* options on the expert form. Likewise, at the end of each section, another comment box was available to write additional comments if there were any (Appendix C).

As previously stated, though the items in the main part of the questionnaire were identical in both instructor and student versions, the students took the Turkish version of it. In order to compare and reconcile the differences between two versions, the English version was translated into students' mother tongue, Turkish, and back-translated into English by an experienced EFL instructor who holds a PhD. There were no meaning differences detected between the Turkish and English versions to reconcile.

After getting the expert opinions, the piloting was conducted with 21 teachers and 21 students. The teachers were teaching at AUSFL at the time of the implementation. Likewise, the students were studying at AUSFL, as well. Also, the participants who involved in the piloting phase, did not take part in the actual data collection process. The questionnaires were sent via Google Forms to all stakeholders. Additionally, Cronbach's alpha coefficient score was calculated for the internal reliability for both the instructor and student versions of the questionnaire. The Cronbach's coefficient was calculated as .86 for instructors and .91 for students, which demonstrates a high reliability for both (See tables 3.4 & 3.5 below). The online version of the questionnaires (English and Turkish versions) using Google Forms were sent to the participants so that the online tool was also tested. The mean scores for the questionnaires were analyzed.

Table 3.4

Reliability Statistics of Piloting with Teachers

Cronbach's Alpha		No of Items	
.860		38	
		N	%
Cases	Valid	19	90.5
	Excluded ^a	2	9.5
	Total	21	100.0

a. Listwise deletion based on all variables in the procedure.

Table 3.5*Reliability Statistics of Piloting with Students*

Cronbach's Alpha		No of Items	
.915		38	
		N	%
Cases	Valid	21	100.0
	Excluded ^a	0	.0
	Total	21	100.0

a. Listwise deletion based on all variables in the procedure.

3.5.2 Semi-structured interviews

To achieve a deeper understanding of the factors affecting participants' test scores, semi-structured interviews were conducted with the instructors, test designers, the deputy principles and students. Heighem and Crocker (2009) values interviews because, when properly implemented, deeper insights into participants' perceptions, experiences and motivations can be achieved. Heighem and Crocker (2009) also explain semi-structured interviews as a way in which "the interviewer has a clear picture of the topics that need to be covered", but at the same time gives flexibility for unexpected directions which touches on new areas.

The interview questions in this study were developed with the guidelines proposed by Ritchie and Lewis (2003), who suggest questions that inspire fulsome responses such as content mapping and content mining questions to obtain breadth and depth are preferable than asking yes/no questions. Bearing these in mind, the questions were developed and then revised after obtaining eight expert opinions who are also experienced ELT professors (see Appendix D for expert opinion and Appendix E for the interview questions).

The interviews were conducted in one-on-one mode, "in which the researcher asks questions to and records the answers from only one participant in the study at a time" (Creswell & Guetterman, 2019, p. 218). Heighem and Crocker (2009) suggest answering some questions before setting up the interview: which are *Who?*, *When?*, *Where?*, *How long?*, and *Under what conditions?* so that the interviewer can allocate enough time for each interview and enough time between interviews to be able to reflect, recover, and plan. Similarly, while conducting the open-ended interview questions, several general steps proposed by Creswell and Guetterman (2019) were followed. First of all, the

interviewees and the type of the interview were identified. Second, a suitable and quiet room for interviews was determined. Next, prior to the interviews, the consent from the interviewees was obtained. Then the questions and the responses were audio-recorded for later transcription. The researcher took some short notes during the interviews when necessary. Since semi-structured interviews were conducted, the researcher was flexible to some extent, and used probes so as to attain more insights from the participants.

3.5.3 Video-stimulated recall protocols

Similarly, the video stimulated recall protocols were conducted with the teachers who had administered the exams before the VSRPs and their students who took the exam in order to triangulate the data. VSR interviews is a powerful tool in which video is used to prompt reflection. It helps reveal participants' underlying cognitive processes during tasks. In this technique, the participants are asked to verbalize their opinions, feelings, and things that shape their decision-making while watching previously captured videos of their performance.

VSRPs are implemented in language teaching and research for various purposes. Initially, it helps learners reflect on the process by reviewing their own language use and behaviors in real-time. Hence, it enables them to be more aware of their strengths and weaknesses in addition to the areas needing improvement in terms of speaking skills. Moreover, learners can notice certain linguistic features and the nonverbal cues by replaying the video. Finally, they can assess their own performance according to the criteria used (Gass & Mackey, 2000). In a similar vein, Lyle (2003) finds VSR interviews as powerful qualitative data tools, especially in teaching and learning since they can illuminate teacher cognition and decision-making. Similar to the interviews, these sessions were also audio-recorded. The VSR interview questions were subjected to a similar process in such after the questions were expert-reviewed and rewritten by the researcher after initial development of the questions (see Appendix F for expert opinions and Appendix G for VSR interview questions).

In the light of these considerations, the VSR interview setting was arranged attentively. Similar stages in the semi-structured interview were applied for the VSR interviews. All participants watched their video recordings individually. The teachers were asked to assess their own assessment, and the learners were asked to comment on their

performance accompanied by the researcher's questions. All sessions were conducted within three weeks after the grades were announced to the learners.

Table 3.6

Time Course for the Data Collection

	Questionnaire	Interviews	Video stimulated recall
Instructors	At the beginning of 23-24 Fall	In the middle of 23-24 Fall	One week after 23-24 Spring Midterm Exam
Test designers	At the beginning of 23-24 Fall	In the middle of 23-24 Fall	-
Students	After 23-24 Spring Midterm Exam	Two weeks after 23-24 Spring Midterm Exam	Two weeks after 23-24 Spring Midterm Exam

3.6 Data Analysis

The population of this study consisted of individuals at AUSFL. The sample included a diverse group of stakeholders involved in EFL education, specifically EFL students, instructors, test designers, and administrators. This purposeful sampling strategy was employed to ensure that data were gathered from participants with varying roles and perspectives within the institutional setting. Also, the inclusion of diverse stakeholder groups allowed for a more comprehensive understanding of the phenomenon under investigation, since each group contributes to the language education process in a unique way. Participants were chosen based on their direct involvement in either teaching, learning, or assessing English language skills, which makes them suitable for giving informed responses to the research instruments.

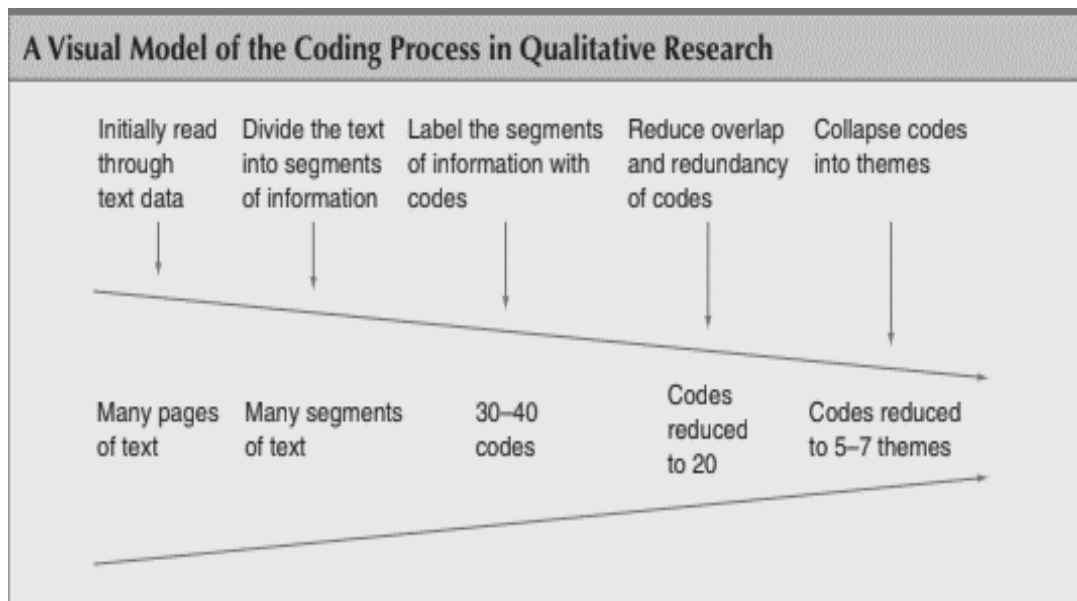
To be able to answer the research question and make an in-depth analysis of the phenomenon, the data were gathered through both quantitative and qualitative instruments. The quantitative data obtained from the questionnaire were analyzed using the Statistical Package for the Social Sciences (SPSS) (Field, 2018). The questionnaire was designed to collect structured and standardized responses, primarily through 4-point Likert-scale items, which enabled the measurement of participants' attitudes, perceptions, and levels of agreement regarding the focus of the study (Joshi et al., 2015). Descriptive statistics, particularly mean scores, were calculated for each factor included in the questionnaire in order to address the research question effectively (Creswell & Creswell, 2018). These mean scores provided a general overview of participants' responses and

were used to identify trends and patterns within the data set (Pallant, 2020). The mean scores of all participants were then interpreted to gain insight into the overall tendencies and perceptions in the responses.

The qualitative data gathered from the interviews took approximately 300 minutes with instructors and 100 minutes with students, and the video-stimulated recall interviews took almost 200 minutes with instructors, approximately 20 minutes for each participant, and approximately 200 minutes with students, around 15 minutes per student. The data were analyzed thematically based on Creswell's content analysis scheme (2012), and his model of coding process was utilized (See Figure 3.4). Initially, the data were transcribed verbatim, and member-checking was conducted together with each participant to ensure the completeness and accuracy of the transcriptions (Lincoln & Guba, 1985). Subsequently, the researcher engaged in repeated readings of the transcripts thoroughly to be able to familiarize herself with the data and started coding the similar chunks. Later, the recurrent codes were determined and labeled by color-coding techniques so that certain themes that emerged based on code-labeling were defined (Nowell et al., 2017) (See Appendix H). To assure interrater reliability, a colleague independently coded more than 25% of the interview transcriptions, chosen randomly, and similar codes were identified. Out of 54 codes, 43 codes were found to be consistent between co-raters and nine of them were revised to enhance coding consistency. Interrater agreement is a noteworthy strategy to provide the credibility and dependability of qualitative findings (Miles, et al., 2014). In order to further assess the trustworthiness of the study, an external audit, who is an experienced EFL researcher, academician and the supervisor of the researcher of this study, critically reviewed the data analysis process to examine the accuracy, correctness and reliability of the data analysis throughout the process. Based on the feedback received, certain coding revisions were done across the data gathered from the face-to-face interviews and the video stimulated recall interviews.

Figure 3.4

Creswell's Model of Coding Process



Reference: Creswell, 2012, p. 244.

4 FINDINGS

4.1 Quantitative Findings

In the quantitative phase of the study, data were collected via a researcher-developed questionnaire administered in two versions: one for instructors and one for students. Both versions employed a 4-point Likert scale, which allowed respondents to express their level of agreement to the statements about students' oral performance and assessment practices. The following sections present the results of the questionnaire responses provided by instructors and students, respectively.

4.1.1 Instructor questionnaire results

To discern the factors influencing English as a Foreign Language (EFL) learners' speaking exam scores, an analysis of instructors' questionnaire results was conducted using descriptive statistics (N=98). Table 4.1, presented below, showcases the mean scores of the questionnaire results, arranged in descending order from the highest to the lowest.

Table 4.1

Instructors' Overall Questionnaire Results

	N	Mean	Std. Deviation
Vocabulary Range	98	2.81	.397
Motivation to speak	98	2.78	.419
Vocabulary Appropriacy	98	2.76	.432
Topic Knowledge	98	2.74	.483
Self-confidence	98	2.74	.461
Fear of Making Mistakes	98	2.70	.522
WTC	98	2.63	.505
Commitment Outside the Class	98	2.60	.605
Grammar Accuracy	98	2.58	.536
Feeling embarrassed	98	2.57	.574
Grammar Range	98	2.55	.520
Task Quality	98	2.55	.594
Task Difficulty	98	2.54	.612
Active Participation in Class	98	2.53	.629
Use of Communicative Strategies	98	2.52	.596
Rater Support	98	2.48	.630
Rater Expertise	98	2.38	.753
Fluency	98	2.37	.505
Following the Criteria/Rubric	98	2.32	.781
Preparation	98	2.30	.692
Rater Cognition	98	2.29	.760
Task Type	98	2.29	.556
Listening Ability	98	2.26	.737

Table 4.1 (continued)

<i>Instructors' Overall Questionnaire Results</i>			
Pronunciation	98	2.23	.552
Use of Warm-up/Ice-breakers	98	2.22	.711
Criteria/Rubric Used	98	2.21	.677
Fear of not being understood by raters	98	2.13	.698
Variety of Tasks	98	2.07	.596
Rater Characteristics	98	1.99	.725
Number of Tasks	98	1.93	.662
Fear of not Being Understood by partner(s)	98	1.78	.780
Duration	98	1.77	.701
Partner(s) in the exam	98	1.74	.663
Being Video recorded	98	1.71	.760
Students' Familiarity with the Criteria/rubric	98	1.65	.801
Rooms' Physical Conditions	98	1.52	.692
Exam Slot Time	98	1.44	.719
Order of Performance	98	1.20	.812

As can be seen in the table above, the instructors regarded *vocabulary range* as the most influencing factor among all factors ($M=2.81$), which is followed by *motivation to speak* ($M=2.78$) and *appropriate use of vocabulary* ($M=2.76$), respectively. On the other hand, *order of performance* ($M=1.20$), which means the order in which students answer the test questions during the exam slot, may not affect the learners' exam scores much, which is followed by *exam slot time* ($M=1.44$) such as the 1st slot in the morning, or the slot after the break or the last slot of the exam day, etc. and *exam rooms' physical conditions* ($M=1.52$) regarding lighting, size, ventilation, etc., respectively.

The results were also analyzed in terms of parts, which are rater-based factors, examinee-based factors including affective, linguistic and other factors, task-related factors, criteria/rubric-based factors and physical conditions. Table 4.2 displays the mean scores of these separate parts in the questionnaire.

Table 4.2

Mean Scores of the Parts in the Instructors' Questionnaire

	N	Mean	Std. Deviation
Rater-based Factors	98	2.27	.740
Examinee-based Factors: Affective Factors	98	2.27	.716
Examinee-based Factors: Linguistic Factors	98	2.55	.540
Examinee-based Factors: Other Examinee-related	98	2.07	.845
Task-related Factors	98	2.35	.650
Criteria/Rubric-related Factors	98	1.93	.791
Physical Conditions	98	1.60	.728

An analysis of the results presented in the table indicates that linguistic factors received the highest mean score ($M = 2.55$), suggesting that participants perceived them as the most influential in the factor that affect EFL learners' speaking exam scores. These were followed by task-related factors, which had a mean score of 2.35. Rater-based factors and affective factors both yielded equal mean scores ($M = 2.27$), positioning them jointly as the third most influential elements. Other examinee-related factors produced a slightly lower mean score of 2.07. Criteria/rubric-related factors followed with a mean score of 1.93, while physical conditions were perceived as the least influential, with a mean score of 1.60.

4.1.2 Student questionnaire results

In addition to the analysis of instructors' questionnaire, learners were also given a similar questionnaire, which was analyzed through descriptive statistics, too. Table 4.3 represents the mean scores of the questionnaire in descending order.

Table 4.3

Students' Overall Questionnaire Results

	N	Mean	Std. Deviation
Number of Tasks	210	2.35	.836
Task Type	210	2.34	.792
Variety of Tasks	210	2.28	.751
Feeling embarrassed	210	2.21	.920
Criteria/Rubric Used	210	2.16	.738
Topic Knowledge	210	2.14	.806
Task Difficulty	210	2.13	.775
Rater Cognition	210	2.12	.955
Self-confidence	210	2.11	.989
Rater Expertise	210	2.11	.903
Students' Familiarity	210	2.11	.790
Vocabulary Range	210	2.10	.735
Grammar Accuracy	210	2.10	.803
Fear of not being understood by the rater(s)	210	2.10	.920
Rater Characteristics	210	2.10	.920
Grammar Range	210	2.09	.851
Use of Communicative Strategies	210	2.09	.851
Fear of Making Mistakes	210	2.09	.908
Being Video-recorded	210	2.08	.832
Motivation to speak	210	2.06	.908
WTC	210	2.06	1.020
Listening Ability	210	2.05	.964
Preparation	210	2.04	.838
Vocabulary Appropriacy	210	2.04	.866
Following the Criteria/Rubric	210	2.04	.874

Table 4.3 (continued)*Students' Overall Questionnaire Results*

Order of Performance	210	1.96	.860
Duration	210	1.94	.851
Use of Warmup	210	1.92	.832
Commitment Outside	210	1.91	.820
Rooms' Physical Conditions	210	1.90	.925
Rater Support	210	1.89	.808
Fluency	210	1.88	.870
Task Quality	210	1.84	.820
Pronunciation	210	1.81	.927
Active Participation in Class	210	1.77	.937
Exam Slot Time	210	1.77	1.016
Fear of not Being Understood by the partner(s)	210	1.76	.988
Partner(s) in the exam	210	1.75	.910

As can be seen, 210 students participated in the questionnaire, and *number of tasks* ($M=2.35$) and *task type* ($M=2.34$) were believed to be the most influential factors for students when they perform a speaking task, which is followed by *variety of tasks* ($M=2.28$). That is, task-related factors affect students' oral exam performances the most from the students' perspective. Besides, *partners in the exam* has the slightest effect on these learners' performances ($M=1.75$). Yet, upon careful examination of the results, it is evident that some items have equal mean scores or exhibit slight differences. To be more specific, although *self-confidence*, *rater expertise* and *students' familiarity* all share identical mean values ($M=2.11$), there are minor variances when compared to *rater cognition*, *task difficulty*, *topic knowledge* and *criteria/rubric used*, with the mean scores of ($M=2.12$), ($M=2.13$), ($M=2.14$) and ($M=2.16$), respectively.

Similar to instructors' results, the parts in students' questionnaire were also examined separately. These parts are rater-based factors, examinee-based factors including affective, linguistic and other factors, task-based factors, criteria/rubric-based factors and physical conditions. Tables 4.4 demonstrates the mean scores of these separate parts in the questionnaire.

Table 4.4*Mean Scores of the Parts in the Students' Questionnaire*

	N	Mean	Std. Deviation
Rater-based Factors	210	2.03	.886
Examinee-based Factors: Affective Factors	210	2.05	.958
Examinee-based Factors: Linguistic Factors	210	2.01	.850
Examinee-based Factors: Other Examinee-related	210	1.91	.895
Task-related Factors	210	2.18	.814
Criteria/Rubric-related Factors	210	2.13	.763
Physical Conditions	210	1.92	.913

When the table above is analyzed, despite small differences among groups, *task-related factors* have the highest scores ($M=2.18$), and is followed by *criteria/rubric-related factors* ($M=2.13$). These were followed by *affective factors*, which had a mean score of 2.05. Rater-based factors yielded a very close mean scores ($M = 2.03$) to affective factors, positioning it jointly as the fourth most influential elements. Linguistic factors produced a slightly lower mean score of 2.01. Physical conditions followed with a mean score of 1.92, while other-examinee-related factors were perceived as the least influential, with a mean score of 1.91. These results reflect students' perceptions of the impact of various factors.

4.2 Qualitative Findings

The qualitative data were gathered through semi-structured interviews and video-stimulated recall interviews conducted with instructors who had evaluated students' oral performances, and the students who took the exams. The semi-structured interviews allowed for in-depth exploration of instructors' perspectives on assessment criteria, grading consistency, and challenges faced during evaluation. In the video-stimulated recall interviews, participants were shown students' oral performances in which they were the assessors and asked to reflect on their scoring decisions and thought processes. This approach enabled a deeper understanding of the rationale behind instructors' judgments related to assessment practices, rater perceptions, and contextual influences, as well as and factors affecting students' performance. The qualitative findings complemented the quantitative data by providing rich, explanatory details. The following sections present the interview results of instructors, interview results of students, video-stimulated recall (VSR) interviews with instructors and VSR interviews with students, respectively.

4.2.1 Interviews with the instructors

As both the test designers and deputy principals are also instructors at AUSFL, in addition to the teachers who directly conduct classes, the interview results for each group, namely teachers, test designers, and administrators, were examined in tandem and discussed under the heading of “Interview Results of the Instructors”. The semi-structured interviews were carried out with 15 instructors teaching at five different language proficiency levels, five test designers from each proficiency level and two administrators. Instructors were selected through random sampling, and all of the instructors were participated on voluntary basis.

Three instructors from each of five different language proficiency levels, five test designers from each language levels and two deputy directors, making a total of 22 instructors, volunteered to participate in the semi-structured interviews which took, on average, 15 minutes for each participant, and approximately 300 minutes in total. The table below highlights the main themes and the codes regarding the factors that have an impact on learners’ speaking performances in their midterm and final exams from the teachers’ perspectives. Also, the excerpts given throughout the section are labeled as I1 (Instructor 1), I2 (Instructor 2), and so forth. Since the interviews were conducted in the instructors’ native language, the translations into English were carried out by the researcher.

The following part explains the factors that affect EFL learners' speaking performances during exams, as observed by the EFL instructors. These factors were categorized under several themes that reflected both internal and external influences on students' ability to perform well in speaking tasks.

4.2.1.1 *Factors that affect students’ speaking performance*

Table 4.5 summarizes the factors that affect EFL learners' speaking performances during exams, as observed by instructors. These factors are categorized under three main themes which are *examinee-based*, *rater-based* and *context-based* factors, each of which highlights different influences on students' ability to perform effectively in speaking tasks. Examinee-based factors encompass individuals’ internal influences which may directly affect performance. Rater-based factors relate to the evaluators' behaviors and perceptions, which may impact scoring. Context-based factors involve external

conditions, which can either facilitate or hinder students' oral performance. By examining these factors, educators can gain a deeper understanding of the factors students face and explore ways to create more supportive and effective speaking assessment environments.

Table 4.5

Factors That Affect Students' Speaking Performance from Instructors' Point of View

Themes	Codes	Positive/ Negative	Frequency	Description
Examinee-based	Affective Factors			Examinee-based factors refer to aspects related to the individual test taker, including their emotions, readiness, linguistic abilities, content production skills, and their characteristics.
	- Anxiety	-	16	
	- Fear of making mistakes	-	1	
	Readiness			
	- Practice	+	16	
	Linguistic Factors			
	- Language proficiency	+/-	10	
	- Comprehension Issues	-	2	
	Content Production			
	- Limited content production	-	5	
Rater-based	- Inability to generate ideas	-	2	Rater-based factors are associated with the evaluators' behaviors and familiarity during the assessment process.
	- Focusing too much on personal experience	-	1	
	Characteristics			
	- being introvert	-	1	
	- willingness	+	1	
Context-based	Behavior			Contextual factors encompass external conditions surrounding the exam that may influence the examinee's performance.
	- tone of voice	+/-	5	
	- speech	+/-	2	
	Familiarity	+/-	2	
	Physical context			
	- Camera	-	3	
	- Pair/group	+/-	3	
	- Waiting (room-setting-cond.)			
	Exam-related context			
	- Task questions	-	4	
	- Evaluation criteria	+/-	2	

4.2.1.1.1 Examinee-based Factors

One of the key themes that emerged is the category of *examinee-based factors*, which encompasses internal influences that may directly impact students' performance during speaking exams. From instructors' point of view, *affective factors* play a particularly major role in shaping students' ability to communicate effectively. Among these, anxiety

stands out as one of the most detrimental impacts, which is often triggered by the pressure of being tested, or social stress. This heightened stress can interfere with students' ability to explain their thoughts precisely and respond with confidence. Moreover, the fear of making mistakes further intensifies anxiety, leading some students to hesitate, limit their speech, or struggle to express themselves fluently.

...The most important factor is the anxiety (factor). Most students who have not taken a speaking test in their previous learning experience and have no experience of public speaking or a speaking test, their anxiety levels can be really high during the exam. I think speaking anxiety is the biggest failure or the factor that negatively affects success in speaking exams. (I6)

...Well, of course, we can call it exam anxiety, exam anxiety because they get very nervous during the exam, I mean, their grades came out well, but in general, what I observed one-on-one is that students get very excited in the exams that I was a jury, especially in the first midterms and if they are taking the exam for the first time, if they are taking the speaking exam for the first time, they have an anxiety problem. (I3)

...That is what we call anxiety, that is, they are very nervous, that is, they cannot speak, they cannot do it. There are also students who know but cannot do it. They really can't speak there when they can speak in a comfortable environment. (I7)

... I think the first reason is that the anxiety level of the students is very high, they said that they were overly excited in the speaking test, they said that although they knew it, they could not perceive it at that moment, they were locked up and they were very excited. (I20)

Besides, *examinees' readiness*, focusing on the level of preparation students have prior to the exam, was found to be another important element. Majority of the instructors identified insufficient practice as an important factor, and stated that learners did not engage in both in- and out-of-class practice adequately. Not knowing how to prepare for the exam and lack of one-on-one speaking practice in the classroom were mentioned by some teachers, which indicates that students' oral performance is likely to suffer.

... Also, we can't do a lot of practice like this in the classroom. I mean, we can't do it with all of them one by one. Before the exam, we showed a mock exam video with you, you know, they became familiar with it, but for example, we couldn't practice with each of them in that way. (I12)

... They cannot spend enough time outside the classroom and internalize the practice of speaking or writing on these issues, the structures they use. (I13)

Linguistic factors including language proficiency and comprehension issues was regarded as another influential factor. These factors highlight certain difficulties mainly related to topic knowledge, vocabulary, and language proficiency. Undoubtedly, students who lack a broad vocabulary or sufficient language skills including where and how to use certain linguistic expressions are likely to struggle in order to understand or respond to exam questions, whereas students with adequate linguistic competencies can reflect better outcomes in their oral achievement tests.

... First of all, the lack of knowledge is the first one, that is, the questions are already from the speaking topics covered in the class. I don't think that 90% of the students who participate in the speaking activities in the class will have a situation like not answering it. This is such a lack of knowledge. Secondly, the second factor is that they don't know the vocabulary, for example, the biggest problem is that they don't know the vocabulary, if they don't know one of them, they can't answer the question. (I7)

... For example, the student hears that word for the first time, which is actually taught, but he does not understand the question either because of the pronunciation or because he does not care much about that word, for example, not understanding the question is very important because sometimes he cannot speak anything, sometimes he cannot understand 2 questions in exams where I ask 2 questions, sometimes the student may not understand 2 questions either because of the word or because he cannot think of anything, so the quality of the questions can also be included in this, of course. (I4)

Additionally, *Content Production*, points to students' struggles with expressing themselves effectively. Specifically, the challenges with content production are regarded

as an influential factor since they lead to the limitation of content development. This is followed by students' inability to generate ideas, which affects the content production negatively. Beside limited content and idea production, focusing too much on personal experiences and questions that address the emotional aspect can limit the depth of examinees' responses.

... During the speech, of course, they also need to produce content, so he needs to have some social intelligence, or is it knowledge, or is it general knowledge. (I8)

... Also, in general, students, of course it varies between those who follow the course and those who don't, but they say that they can't come up with ideas about the questions, that is, they say that I can think of one thing and the rest is nothing. (I11)

Finally, *learners' characteristics* is considered as another important factor among examinee-based factors. Learners' introversion and willingness can have both positive and negative effects on their exam performance. Introverted learners might struggle in speaking tasks in the exams as they demand active participation and social interaction. In a similar vein, a high level of willingness can enhance motivation and engagement, leading to better outcomes, while a lack of willingness may result in decreased effort and lower performance.

First and foremost is the level of readiness or willingness of the students. Generally Turkish students are a bit reluctant to speak. (I14)

In brief, these emotional, characteristics and linguistic barriers as well as students' readiness and content production skills highlight the importance of creating a supportive and low-stress testing environment to help students perform at their best.

4.2.1.1.2 Rater-based factors

Another crucial theme is *Rater-based Factors*, which basically includes teachers' behavior and familiarity, which can both affect examinees' performance positively or negatively. Namely, when raters exhibit a positive demeanor during the exam by using a soft tone of voice and clear speech, students' comfort levels may increase, potentially leading to improved performance and reduced anxiety. Likewise, examinees' familiarity

with the examiners is stated as having either positive or negative effects. In a general sense, an unfamiliar rater can increase students' anxiety and disrupt their performance, whereas familiar examiners can create a more comfortable testing environment, helping students feel at ease and perform more confidently.

... The demeanor of the teachers can be important. Here, the lecturer should be more understanding and give feedback by giving various clues. (I14)

... The attitude of the examiners can be influential. It can be positive or negative, of course. I mean, in general, students are very excited, especially in speaking exams, and they express this after the exam, for example, the teachers who gave the exam approached us very well. They were very friendly to us. When this happens, they say it relaxes them. It can easily demoralize children, or it can easily demoralize them for the second and later questions about a question they said wrong. But it crossed my mind that maybe more encouraging attitudes or something like that could help them to continue the exam. (I10)

4.2.1.1.3 Context-based factors

Context-based Factors can influence performances, as well. The *physical and social exam* conditions such as being video-recorded and the imperativeness of working in pairs or groups of three and the conditions of the waiting room (where students wait prior to the exam) were seen as influential elements. The existence of the camera was deemed as a negative influence, whereas being paired up or implementing the tasks in a trio can have both positive and negative effects, depending on the dynamics of the individuals. Furthermore, an instructor highlighted the negative effect made by students who had taken the exam prior to them.

... There is a camera recording. I mean, we told them that there will be a camera recording during the exam, but of course we don't do such a thing during the lesson. So, when they see the camera in the exam, they are a little bit scared. (I12)

... They say the pairs they take the exam with affected them. For example, my students said "my friend next to me didn't intend to talk too much, he didn't care too much, it affected me too". (I15)

Exam-related context, involving task questions and the evaluation criteria, is another context-based factor. Four instructors stated the positive effects of the questions matching with the course content, which helped students produce better responses to the questions. Students can achieve successful results in these exams since the questions are aligned with their study materials and learning outcomes as these are achievement tests. Finally, *evaluation criteria* reflects the concerns related to the clarity and appropriateness of descriptors in the evaluation criteria. As observed by two instructors, if learners are unclear about the exam criteria and thus, how their responses will be evaluated, they might face challenges. Also, one instructor stated the descriptors push him up to give higher grades.

... In a positive sense, they are familiar with the topics that we talk about in the course content, that is, they are familiar with the topics because questions are asked on the topics in the book. (I11)

... They (Students) don't know the criteria very much. We show them, of course, but they don't say, "I wonder what they are evaluating me according to." Unfortunately, they think that we only look at grammar. So, they focus a lot on grammar. I think they are a little bit uninformed about the criteria. (I12)

...For example, the criterion... the criterion that we use, I'm doing a speaking exam after all, and sometimes, you know, the subject content is very good. He (a student) tries to talk about it, but it's like I'm actually commenting a little bit. You know, I understand him, but when it comes to grading, my grades go up with the criterion, you know, those descriptors push me up for some reason, so that also affects the grades. (I9)

In sum, Table 4.5 provides an overview of the complex factors that affect speaking performance in exams, where examinee-, rater- and context-based factors all play a significant role in shaping students' success. Examinee-based factors mainly include affective factors, students' readiness, linguistic factors, learners' ability to produce content and their characteristics. Rater-based factors involve rater's behavior including how they use their tone of voice and speech, while context-based factors are more related to physical context such as camera existence and pair/group tasks, and the exam-related context such as task questions and the evaluation criteria. By examining these factors, examiners can gain a deeper understanding of the positive factors and the challenges

students face and explore ways to establish more supportive and effective speaking assessment environments.

4.2.1.2 Challenges that affect students' speaking performance

Table 4.6 below provides an overview of the challenges that students face during speaking exams from the instructors' perspective. These challenges can be categorized into three main areas: *examinee-based*, *rater-based*, and *context-based challenges*. Examinee-based challenges stem from factors related to the individual test takers. Rater-based challenges arise from the evaluators' behaviors and perceptions. Context-based challenges, on the other hand, involve external conditions that may impact performance. Understanding these difficulties can provide valuable insights into the challenges students encounter and may help improve speaking exam practices.

Table 4.6

Challenges that affect students' speaking performance from instructors' point of view

Themes	Codes	Frequency	Description
Examinee-based	Affective Factors		Students' struggle due to examinee-related stress and anxiety, readiness, linguistic competencies and negatively impacted their performance.
	- Anxiety	17	
	- Fear of public speaking	3	
	- Fatigue	3	
	- Issues on the exam day	2	
	- Negative experiences	1	
	Readiness		
	- Practice	8	
	Linguistic Factors		
	- Lack of vocabulary	6	
Rater- based	- Pronunciation	4	Jury behavior including their attitude, tone of voice, being high/low graders, and familiarity with the examiners had an impact on performance.
	- Grammar	4	
	Content Production		
	- Inability to generate ideas	10	
Context-based	- Comprehension issues	4	Challenges related to exam setup such as recording, and exam tasks, peers, format and the criteria affected students' performance.
	Behavior	8	
	Familiarity	5	
	Physical context		
	- Video recording	6	
	- Environment	3	
	Exam-related context		
	- Tasks/Questions	5	
	- Peers	5	
	- Format	3	
	- Criteria	1	

4.2.1.2.1 Examinee-based challenges

Similar to the overall factors mentioned above, examinee-based challenges involve *affective factors*, *content production*, *examinees' readiness*, *linguistic factors*, and *comprehension issues*. Regarding affective factors, anxiety was deemed as a major hurdle for many students in oral tests, which was followed by fear of making mistakes, fatigue, issues on the exam day and negative previous experiences. Instructors believed that students considerably struggle due to exam-related or social stress, psychological pressure such as fear of public speaking, or issues on or before the exam day, which can exacerbate students' nerves.

... But if you say the biggest difficulty, I think anxiety, that is, the anxiety of the exam, because even if the student is in a position to speak at that moment, he can forget, so let me answer like this, you know, the biggest difficulty is anxiety. (I4)

In terms of difficulty, it is actually a little bit psychological again, but I think the biggest challenge they face is that they have not suppressed that anxiety. (I5)

... I think that students' speaking anxiety in general, that is, their anxiety about speaking in a foreign language, lies in the fact that no matter what kind of exam we do, students generally perform lower than their normal levels due to both the anxiety brought about by the exam and their high anxiety about speaking in a foreign language because it is during the exam. So I can actually say speaking anxiety. (I18)

In other words, even something that happened on the way to the exam hall can affect the student. He might have woken up late and run to the exam or take a taxi. This will definitely affect The performance of the exam in the exam. It is not specific to that day, but the mood he was in at that time, a bad experience he had will affect him negatively. A physical discomfort can definitely be effective during the exam. Apart from that, fatigue can affect. (I6)

... or other factors, other than the exam, can affect students' success in the exam, such as discomfort, etc. (I22)

Content production difficulties were also mentioned by a great deal of instructors. Particularly, instructors highlighted the learners' inability to generate ideas during the exams. Likewise, difficulty in understanding task questions and the instructions point to a lack of clarity in responding to prompts. That is, since they struggle to come up with ideas, students also endeavor with organizing their speech which prevents them from delivering more coherent and comprehensive responses.

... I think the biggest challenge for students is to generate ideas rather than language. Well, when the student answers yes or no, they don't think about it as if it is solved. Why is that? Normally, when we speak in Turkish, I don't know, do you like reading books? I do, what kind of books do you read? That conversation evolves in some way. But students don't have that natural thinking, I mean, they look at it as if it's another language, as if I gave a complete answer and it's over. I think this is the biggest problem. (I17)

Examinees' readiness also affected the students' exam performance negatively. The impact of insufficient preparation and practice, both inside and outside the classroom, caused to be a hurdle for them. The lack of practice as well as the lack of knowing how to get prepared for the exam reflected students' being unsure about their expectations for the exam. Besides, not paying enough attention to in-class activities was seen as a challenge for the examinees. Also, in the course of the exam, some instructors believed that some students hurried to answer the question without using the preparation time for the task efficiently.

The biggest challenge. That is, the student is not ready enough. That is what I observe. (I3)

... yes, I mean actually not practicing enough. In the classroom, yes, we are trying to prepare a very communicative syllabus. We are also trying to implement this, but this is what we call learner, that is, the situation where students prefer to remain silent in the classroom. (I6)

Additionally, *linguistic factors* created certain challenges since they comprise issues related to students' language proficiency. As mentioned by six instructors, challenges in vocabulary created struggles in expressing the ideas. Similarly, pronunciation difficulties,

lack of grammatical knowledge and grammar-focused thinking underlined how linguistic limitations impeded students to express themselves explicitly in the course of the exam.

... they don't understand some words in the exam. They don't understand some of the words in the questions. It gives them a big panic. What does this mean, what will happen if you don't understand, will I get zero or something like that? (I12)

... I mean, this is the biggest difficulty, I think the lack of vocabulary is a very big difficulty because it can be even the smallest word or even like this. I mean, he doesn't understand because he doesn't know the correct pronunciation of the word. So you say what or something like that. You repeat it but not with the pronunciation he knows. You pronounce it with the correct pronunciation. That's why he doesn't understand and he goes like that, he says no answer or something like that and finishes. (I7)

4.2.1.2.2 Rater-based challenges

As mentioned in overall factors above, *rater-based challenges* also negatively impacted students' performance. Especially, examiners' negative attitude, harsh tone of voice and being high/low graders had an impact on students' performance. Furthermore, unfamiliar examiners created tension on the test-takers, exacerbating students' nerves, and resulting in poor performance.

The biggest difficulty is having teachers they don't know, first of all they don't feel very safe sometimes. Well, the atmosphere of trust in this class is the feeling of knowing someone and thinking that they can understand themselves, and of course they know that mistakes will be corrected, but there is no such thing in the exam and they will be evaluated by their performance and I have heard from a few students that they do not feel good in the exam, but after all, this is the nature of speaking. (I9)

... a negative attitude of the examiners can affect them. (I10)

... students may have some perceptions on the basis of their professors, like low-grader or high-grader. (I21)

4.2.1.2.3 Context-based challenges

The final challenges stated by the instructors were *context-based challenges*. These challenges included physical context and the exam-related context. Regarding physical context, instructors drew attention to the pressures of video recording and the tense exam environment.

... again, this will be the psychological barrier. Of course, this is the presence of cameras and so on. Especially if they are taking the oral exam for the first time, it can be a problem to get through it. (I13)

... If I think about midterm and final, the exam environment and the tension created by the exam environment is probably the biggest challenge in that sense. Will I understand the question or not in front of the teacher? ... but there is a pressure of being recorded, which they themselves also mentioned. I mean, the stress of the exam is one thing, but I think it is probably the fact that they are recorded and that they take the exam in front of a teacher or 2 teachers they do not know. (I11)

... for example, students say that the camera recording makes them very nervous, some of them, not all of them, but some of them have such an arrest when they see the camera, etc. (I5)

Furthermore, the exam related challenges included the exam tasks, format, the criteria and the peers, which also provoked additional stress. Although the questions were aligned with the topics covered in the lessons, some students still found some questions surprising as if they had not seen them before. Moreover, owing to the format of the exam, examiners' inability to give prompts or clarify meaning by paraphrasing further compounded difficulties which hindered their success. Time limit was deemed as another challenge, and the criteria's being open to interpretation might affect their performance. In addition, exam peers had negative effects on the performance. Specifically, in situations where students were not well-acquainted with each other, or when there was a disparity in language competencies between peers, the spoken performance was negatively impacted.

I would say the biggest challenge is the surprise effect. I mean, what will I get when I take the exam? As a subject, as a question. (I8)

Question format. Exam format. ... You know, the person who writes the question, you know, the person in the testing unit definitely has an outcome in his/her head, something in his/her head, something targeting something. It is always possible that it may or may not make it across (to students). So, it may not pass (to students). (I2)

... and 10 minutes is not always enough. So they may not be able to show their real performance. I think this is also a big obstacle, a challenge for them. I mean in terms of showing performances. (I1)

... may be affected by the friend next to him/her. If he/she is good for him/her or if he/she is worse than him/her, he/she may not be able to focus fully because he/she will ask questions such as if I were to do something like an interaction with him/her, how would it affect my grade positively or negatively, and if he/she is at a high level, the same thing may take the form of how can I do something with him/her this time? I can carry out an event. (I16)

All in all, the complex challenges that affect students' speaking performances, stem from a combination of examinee-based barriers, rater-based issues, and contextual factors, all of which contribute to the difficulties students face during speaking exams. Examinee-based barriers, such as anxiety, linguistic struggles, and limitations in content production can hinder students' ability to communicate effectively. Rater-based issues such as positive behavior and familiarity bias may further affect students' performance. Furthermore, contextual factors like the physical exam setting, camera recording, and the structure of task questions can obstruct students' ability to express themselves confidently. These interconnected challenges highlight the need for a more comprehensive approach to speaking assessment, ensuring that exams accurately measure students' abilities while minimizing unnecessary obstacles.

4.2.1.3 Facilitators that affect students' speaking performance

The factors that instructors perceive as facilitating students' speaking performance were categorized into three main themes, emphasizing the strategies and conditions which can pave the way for an effective and supportive oral assessment experience for EFL learners. As displayed in Table 4.7, these facilitating factors that affect students' speaking

performances in the achievement tests fall into three categories: *examinee-based*, *rater-based* and *context-based* factors. Examinee-based factors include some individual aspects such as students' preparedness, confidence, and prior practice, which enhance their ability to perform well. Rater-based factors involve fair and consistent evaluation practices, constructive feedback, and a examiners' supportive demeanor, all of which may positively influence students' oral exam performance. Context-based factors refer to external conditions such as a comfortable testing environment, clear task instructions, and opportunities for interaction, which help reduce stress and encourage effective communication. These factors also include peer effect since these test takers were assigned into pairs prior to the exam by the testing office members, and were required to take the exam in pairs. However, it should be noted that taking the exam individually was not permitted. Therefore, in cases where a student failed to appear, the testing unit members reassigned the remaining student to the following pair, resulting in a group of three test takers.

Understanding these facilitating factors enables test evaluators to refine their evaluation practices, particularly in oral testing contexts, ensuring that oral tests provide students with an opportunity to demonstrate their language competencies in a comprehensive and meaningful manner.

Table 4.7

Facilitators that affect students' speaking performances from instructors' point of view

Theme	Codes	Frequency	Description
Examinee-based	Achievement-oriented exam	12	Students feel more comfortable and confident when the exam includes familiar topics or when they practice.
	Practice	8	
	Learning strategies	3	
	Familiarity with the exam procedure (e.g., mock exams, guidelines)	3	
Rater- based	Familiar instructors	7	A positive and supportive attitude from the jury helps reduce anxiety and create a better exam environment.
	Friendly demeanor of instructors (e.g., smiling professors)	4	
	The way examiners speak	1	
	Teachers' maximum use of English in class	1	
Context-based	Peer effect		A well-organized exam environment allows students to focus on their performance
	- Exam with classmates	6	
	- Partner relationship	2	
	- Taking the exam in pairs	1	
	- Ideas from partners	1	

Table 4.7 (continued)

Facilitators that affect students' speaking performances from instructors' point of view

- Being at a similar level with the pair	1	rather than logistical distractions.
Exam (questions) format	4	
Exam environment		
- Physical environment (seating, lighting, temperature)	3	

4.2.1.3.1 Examinee-based facilitators

One important advantage that affected student performance was the tests' being an *achievement-oriented exam*. Majority of the instructors perceived that students felt more confident and achieved better results in midterm and final exams as these are achievement tests, and thus, the questions were grounded on familiar topics or content that had already covered in the classroom previously. Namely, the students were familiar with the topics; therefore, when they attended classes regularly and were aware of the course objectives and goals, they were likely to perform better.

In these exams, we always ask about the subjects they cover in the book from something we have talked about, studied, read, listened to in the lessons, so this is the biggest convenience since it is an achievement test, so there is no difficulty for a careful student who follows his lessons regularly. (I1)

... the fact that we have already talked about the questions in the lessons asked in the exam, (this) creates a familiarity with the subjects. ... if they came to school on the day it was taught, it is a subject that the student is actually familiar with. So they can talk about familiar topics, they can talk about it or they are held responsible for it. (I4)

Another facilitator was the practice done by the students. A number of instructors mentioned the positive effects of sparing time and effort for extra practice and preparation, not only in but also outside the classroom, boosted their scores. Likewise, the use of appropriate learning strategies such as improving skills in linking ideas, doing timed speaking exercises and recording their speech and listening to it afterwards was regarded as highly useful in oral performance, as it provided students with the opportunity to assess their own performance, identify areas for improvement, and built confidence in

their speaking abilities. Furthermore, three instructors noted the importance of students' familiarity with the exam procedure. For example, students watched mock exams, which is a simulation of the real exam acted out by the teachers in a real-like situation, offered students ideas about the exam format and process. Besides, guidelines were given to the students prior to the exam, which helped to reduce uncertainty, and enhanced familiarity with and readiness for the actual exam.

It would be much better for them if they incorporated speaking English more into their lives, which they never do, we only speak as much as we can for 45 minutes in class. Then they forget it in their social life. I mean when they go home. Therefore, if they could speak a little bit of English at home with their friends or listen to them, of course it would make their work much easier. (I11)

Talking for long periods of time. For example, talking about a topic for one minute without stopping, talking for one and a half minutes. These allow them to work in a timed manner, and they can also talk about the same topic, for example, one minute with a friend, the same question with another friend, the same question with another friend... When they repeat it four times, they may have gained some fluency on that topic and enriched the content when they hear ideas from each other. I think these kinds of activities can be useful. (I14)

4.2.1.3.2 Rater-based facilitators

Another key factor in students' speaking performance is related to the raters in the exams. Specifically, the positive demeanor, attitude, the way examiners speak, and the familiarity with them promote students' success. Seven instructors, for example, emphasized that the presence of familiar professors has a positive impact on mitigating the effects of stress, since students feel more comfortable with the examiners they already know. Moreover, when examiners demonstrate a friendly and encouraging attitude, they are more likely to put students at ease so that they feel less anxious. This positive interaction leads to a fruitful environment for students to perform more satisfactorily. Furthermore, the way the examiners speak during the exam can affect examinees. These involve pronunciation, speed of speech (fast/slow) and the high/low pitch. It is underlined that examiners' having clear pronunciation and effective use of tone of voice reduce nervousness and foster

achievement. Furthermore, students' being exposed to the target language through the teachers' maximum use of English in the classroom was also mentioned as another facilitator.

Of course, our attitude, that is, our openness to be helpful. That is, speaking slowly and with correct pronunciation, showing a smiling face. (I2)

... but for students, the fact that their own teachers are doing the exam is perhaps a factor that makes them a little more comfortable... Because they feel more comfortable when they see their teachers. Maybe they feel a little bit relieved in terms of stress because they see someone they trust and like. (I5)

4.2.1.3.3 Context-based facilitators

Context-based facilitators involve peers, exam format and the physical context during the exam. *Peer effect* was seen as a crucial factor that influences students' speaking performance. Six instructors observed that when students take exams with a familiar partner, they feel more relaxed and confident because the relationship with a partner can reduce feelings of isolation and provide emotional support during the exam. Additionally, positive relationships between the pairs and taking the exam in pairs also allow students to collaborate and feel less intimidated by the exam environment. Examinees' having similar language levels and being able to draw on the ideas shared by the exam partners were also viewed as other facilitators. Namely, a more comfortable atmosphere, in which students are more likely to focus on demonstrating their speaking abilities, can be created by mutual support.

... they usually take the exam with friends they know. In the same class, this way, there is no anxiety caused by an unfamiliar person. (I6)

.. And friends, you know, entering with friends gave them confidence, and even as I just said, one of my students, who I thought would get a high score, got a very low score. I mean, I remember he got 50. They paired him with someone from another class and he said it had a great effect. (I9)

Beside peer effect, *exam format* had a remarkable influence on students' performance. Along with being two parts in the exam, some teachers found picture-prompted questions efficient in generating ideas. Likewise, instructors suggested that asking follow-up

questions are beneficial because they give students the opportunity to respond elaboratively and to display their speaking skills more comprehensively. Additionally, though not allowed, paraphrasing questions was thought as another useful technique since it facilitates students' understanding of the question and give them the confidence to answer accurately.

The things that make it easier for students can be the photos we use in the exams, if we think in terms of the exam format. (I18)

It is good to give them time because I remember in some exams, we didn't give them time at all. Thinking makes it easier for them. (I9)

When I pave the way a little bit for the students with extra "follow-up" questions, I can get better things. I have at least one material from the students that I can evaluate. This makes it a little easier for the students. Paraphrasing questions can sometimes make it easier for students. (I15)

Finally, *exam conditions*, particularly referring to the physical environment of the exam room, plays a vital role in facilitating students' performance, as well. Instructors stated that a well-organized exam environment including seating arrangements, lighting, and room temperature also contribute to students' comfort during the test. In other words, a positive physical environment can lower stress and allow students to perform their best.

... I mean, we actually apply a standard in our exams. I mean, we even tell our teachers how to ask the questions they will ask in the speaking exam. We give a directive about this. Here is how the procedure of meeting the student first will be, then we give everything, even the wording of the question stem of the exam, so what we normally expect from our teachers is to follow this directive. (I21)

We (a test designer) give the students a guideline about what kind of questions will be asked before the exam. Since the lecturers discuss this with the students in their classes, the students are already familiar with the exam procedure. I think this provides a comfort. Because instead of facing something unknown, they go to a process they know. (I19)

... there can be a mutual seating arrangement. You know, a seating arrangement where they can feel comfortable to communicate with the teacher. I don't know, light, temperature, etc. I guess these also affect it. (I10)

In conclusion, instructors highlight a variety of factors that can facilitate students' speaking performance during exams which are categorized under themes: examinee-based, rater-based and context-based facilitators. Among these themes, examinee-based facilitators include achievement-oriented exam, practice, learning strategies and familiarity with the exam procedure. Rater-based facilitators are familiarity, friendly demeanor, the way they speak and the use of English in class. Final theme was gathered around peer effect, exam format and the physical context of the exam. All these elements contribute to creating a comfortable environment where students feel confident and prepared, allowing them to perform to the best of their abilities.

4.2.2 Interviews with the students

In addition to instructors, interviews with three students from each of the five language proficiency levels were conducted with volunteers, and the interviews took approximately 100 minutes in total. The findings of the student interviews were also grouped in three main categories: factors, challenges and facilitators that have an influence on students' speaking performance from their viewpoint. The overall factors that affect the performance were displayed in Table 4.8, while the challenges and the facilitators were represented in Table 4.9 and Table 4.10, respectively. Also, the excerpts from the interviews were presented anonymously in order to conceal identities, and are labeled as S1 (Student 1), S2 (Student 2), and so forth. Since the interviews were conducted in the students' native language, the translations into English were carried out by the researcher.

4.2.2.1 Factors that affect students' speaking performance from students' point of view

When analyzing the factors that affected learners' speaking exam performance from the students' perspective, the themes emerging from their responses were classified into three main categories. Consistent with the findings from the instructors' interviews, these themes are *examinee-based factors*, *rater-based factors*, and *contextual factors*.

Table 4.8*Factors that affect students' speaking performance from students' point of view*

Theme	Codes	Positive/ Negative	Frequency	Description
Examinee-based	Readiness	+		Constant practice, preparation and positive self-perception develop speaking fluency, confidence, and readiness for the exam, whereas anxiety impact students' performance adversely.
	- Practice		9	
	- Vocabulary revision	+	2	
	- In-class participation	+	2	
	Anxiety	-	7	
	Self-perception			
	- Feeling motivated	+	2	
	- Reflecting the language in our lives	+	2	
Rater-based	Behavior	+ / -	4	Supportive and familiar teachers play a vital role on examiners' performance.
	Familiarity	+ / -	3	
Context-based	Peer effect			The quality of exam questions as well as the exam conditions can either create more positive results or affect performance adversely.
	- Familiar partner	+	5	
	Exam questions	+ / -	3	
	- Unannounced questions	-	1	
	Exam conditions			
	- Slot	+ / -	1	
	- Quick process	+	1	
	Exam (physical) context			
	- Video-recording	-	1	

4.2.2.1.1 Examinee-based factors

Examinee-based factors encompass internal factors such as *readiness*, *anxiety* and *self-perception*. Students mentioned the importance of getting ready through constant practice both inside and outside the classroom. Some of them also emphasized the facilitating role of peer-speaking practice as it could enhance oral skills in a collaborative and interactive manner. Besides, being active in in-class activities and tasks as well as doing vocabulary revision were also acknowledged as complementary strategies that helped improve their speaking fluency and, as a result, developed their overall readiness for the exam.

In fact, I think the biggest reason why I got this score is that firstly, I focused on the exam, and secondly, I think that putting the speaking practice in the center of my life and using it in daily life helped me the most. I think that the quality of the education is affected by the way we improve ourselves

depending on the activity we learn in the classroom and how we prepare ourselves before the exams. (S1)

On the positive side, the vocabulary repetition helps me to understand the lesson well and when I do general repetition or when I talk to a foreign friend or someone who speaks English in games, it affects me more. (S9)

Extra practice. Extra practice is to look at the sites I said. It could be a cartoon on the A1-A2 thing (level). It can be a serial. You can look at them and do a little bit of speaking with your friends and there is also something to learn to talk more about the things we normally talk about. (S13)

I think on the other hand, you need to work regularly for speaking, you know, you need to repeat it beforehand. I think I'm a bit lacking in this too. I think I'm also a bit lacking in vocabulary. I think these are important. These are actually important for me. (S15)

However, *anxiety* was regarded as a significant challenge, mentioned by almost half of the participants. Stress and nervousness, as is known, could have detrimental effects on students' performance in the oral exams. These factors not only undermined confidence but also hindered fluency and disrupted students' ability to demonstrate their linguistic competence and communicative skills effectively. Consequently, heightened anxiety might lead to increased errors, hesitation and an overall reduction in performance quality, ultimately influencing their assessment scores.

I think stress affects performance the most because most people cannot perform as well as they normally do in speaking exams. (S5)

From a bad point of view, as I said in the first question, it could be a stress-anxiety problem again. (S9)

What affects me negatively is my own personal state of constant tension and stress. This affects me. (S11)

I mean, I'm not sure, but really stress, yes, I get very nervous. And that's why, even though I understand the question, sometimes I can't really answer it properly. So, I can only say stress for me, really. (S14)

Students' *self-perceptions* of themselves also affect their performance. Feeling motivated, for example, was declared as a significant factor since it enhanced confidence and boosted ability. In a similar vein, feeling motivated supported students' performance, whereas lack of motivation hindered their success which resulted in undesired exam results. Furthermore, the ability to reflect language use in one's daily life was regarded as another valuable complementary skill that could further help to achieve better performance outcomes.

The most important thing is to reflect this in our lives. (S2)

I think motivation is also very important. (S12)

4.2.2.1.2 *Rater-based factors*

Rater-based factors, involving jury's behavior and familiarity, were also identified as other critical factors that were influential in demonstrating students' oral skills. Particularly, a supportive attitude from the jury created a positive and less intimidating atmosphere for examinees, while unfamiliarity with the jury members was seen as an important hurdle. Furthermore, students acknowledged the comforting and less stressful assessment atmosphere in which the examinees were acquainted with the examiners. In other words, when the teachers who administered the exam were familiar, students performed better.

In speaking exams, I think that the attitudes of the teachers, especially their gestures, eye contact and gestures are also very important. If the teacher is comfortable in front of you, you are also comfortable. If the teacher looks you in the eye and gestures in a way that confirms you, you are also relaxed this time because it is a bit of a reactionary environment. (S2)

It can also be the attitudes of the teachers, because sometimes I feel the need to make eye contact with someone at some points, even if you don't make eye contact while talking, and when I can't make that eye contact or when I encounter a gesture style that is harsher, as if something is going negatively, it really stresses me out a lot. (S4)

4.2.2.1.3 Context-based factors

Context-based factors, deemed as other influential factors, involve *exam questions*, *exam context* and *peer effects*. The quality and the clarity of the questions and their comprehension by the students indicated their significance in achieving oral success. When exam questions are clearly formulated and easily comprehensible, especially when they were equipped with related pictures, test-takers were more likely to respond accurately and appropriately, thus revealing a more realistic representation of their actual proficiency level. Additionally, providing the questions in advance was perceived as a facilitator for understanding and getting ready for the exam, whereas the absence of it was seen as a challenge, possibly affecting performance adversely.

If the photos in the question or the sentences in the question are close to each other, we can actually think faster and give answers by thinking more analytically. It is actually difficult to catch a context between photographs that are far away from each other even in our mother tongue. (S2)

... I think the factor that affected me the most was that the question was not given, ... (S8)

Questions... they are the most important (thing) in the first place. (S12)

Exam Conditions were also noted as factor although it was not mentioned by the majority. For example, a student mentioned the importance of the exam slot time, noting that she believed earlier slots were better. One student liked the duration of the exam and stated that it was a quick process, which was probably because it reduced waiting-related anxiety. Conversely, being video-recorded was perceived as negative, potentially adding stress to the exam experience.

In the exams I took in the first semester, the camera was affecting me a lot. I felt like I was under a lot of stress. (S4)

It's also a very fast process. We talk in 15 minutes, we finish. (S5)

I think it is an advantage to take the exam in the early hours because the questions are easier, it becomes more difficult in the later hours. (S8)

Peer effect also played a pivotal role on students' achievement in oral exams. Some students stated that taking the exam with a familiar pair, generally a classmate, reduced

stress and helped them feel at ease. Also, when students had the opportunity to do practice with their peers, they felt relaxed, and thereby produced better results.

It's more comfortable when I enter (take the exam) with a partner. It's more beautiful. So, you can talk more comfortably. And if it's someone you know, you can be more motivated. (S10)

If I trust my friend next to me, you know, if he is someone I know and I trust his speech, he also relaxes me. I mean, it affects my performance, after all, at the discussion (part), for example, if he constantly asks me questions or something, he will save me and save himself, so this relaxes me. (S7)

To sum up, the interviews with the students indicate the complex nature of speaking exams and their effects on students' performance. The majority of the students underscored the importance of constant preparation and practice along with the other elements such as anxiety, jury features exam conditions, self-perception and exam questions that affected EFL learners' speaking proficiency.

4.2.2.2 Challenges that affect students' speaking performance from students' point of view

Although two students (S1 and S3) from the highest language level (A+) stated that they found no difficulty regarding the speaking exams, Table 4.9 presents the challenges other students face during their speaking exams, as perceived from their own perspectives. These challenges were categorized into three key themes, which are *examinee-based*, *rater-based* and *context-based challenges*.

Table 4.9

Challenges that affect students' speaking performances from students' point of view

Theme	Codes	Frequency	Description
Examinee-based	Anxiety	8	Various forms of stress and anxiety as well as linguistics issues impacted students' ability to perform.
	Linguistic factors		
	- Vocabulary issues	5	
Rater- based	Behavior	1	Teachers' unsupportive behavior and vague guidance boosted stress during speaking exams.
	Familiarity	1	
	Accent	1	

Table 4.9 (continued)*Challenges that affect students' speaking performances from students' point of view*

Context-based	Peer effect		Negative peer effects, questions not given in advance, remote reding of the questions and limited time during the exam made it hard for students to think and respond effectively.	
	-	Unfamiliar exam partners		2
	-	Negative comments from previous examinees		2
	-	Pairing with more/less fluent partner(s)		1
	Exam questions			
	-	Difficulty understanding due to remote or unclear format		2
	-	Unprovided written questions		1
	-	New topic		1
	Duration			1
	Exam conditions			
	-	Multiple examiners		1
	-	Video-recording		1

4.2.2.2.1 Examinee-based challenges

Among *examinee-based challenges*, anxiety emerged as the most influential factor, including stress and nervousness. It encompasses various triggers, such as stress in general and the pressure related to exam itself. These factors can adversely affect students' ability to convey their thoughts clearly, and in return, hinder their performance.

I think the most challenging thing is to be under stress because we can all somehow talk in class, whether with our teacher or with our friends. (S5)

I think the most difficult thing for me is that I can't think at that moment either because of stress or because it happens so fast. (S6)

When I found out that we couldn't read the questions, you know, that made me a bit nervous when I first entered. (S7)

... stress, Because for example, someone who is very good at speaking may not be able to speak at all due to stress. (S8)

Besides anxiety, linguistic factors impacted students' performance in a negative way. Specifically, vocabulary-related issues were deemed as a vital challenge that hindered students' ability to demonstrate effective communication during speaking exams. These issues involve gaps in vocabulary knowledge, difficulties with complex vocabulary and

forgetting words during the exam, which caused barriers for students to express ideas clearly.

The vocabulary, I forget the words. Also, as the level increases, the words become more difficult as the level increases. (S9)

... it could be a lack of vocabulary, in fact, for example, I don't know many words... (S10)

4.2.2.2.2 Rater-based challenges

Another theme that emerged in students' interviews was related to rater-based elements. The students noted the effects of examiners' behavior, familiarity with the students and accent as determinants on their success. Namely, the administration of the exam by unfamiliar jury members may cause a sense of discomfort, whereas supportive behavior and clear guidance were acknowledged as elements that ease anxiety and enhance performance. Teachers' accent and attitude are acknowledged as additional concerns that limit students' ability to perform well.

Apart from that, I think the attitude of the teacher in front of us while (we are) answering the questions in the speaking exam also affects us because when people speak, especially we are students and when there is an exam, we want to be approved, you know, am I speaking correctly, am I doing well, am I doing badly? We want to get a reaction, and since it is an exam, of course, our teachers pay a little more attention to this, but I think it should be more approved, that is, even if the other student speaks badly, because he is under stress at that moment and cannot do it. I think it should be ignored a little bit and approached more positively. (S5)

Sometimes the instructors' accent affects me. Since I am usually closer to American English, the accent confuses me a lot when someone speaks British English. (S12)

4.2.2.2.3 Context-based challenges

Thirdly, *context-based challenges* encompass peer effects, exam questions, duration and exam room setting. *Peer effects* during the exam could cause discomfort among students. To illustrate, pairing with unfamiliar or with more or less fluent exam partners could lead

to feelings of anxiety and inadequacy. Moreover, negative comments from students who took the exam previously were seen as another challenge, since they could heighten anxiety and lower confidence.

I felt that I could not match the way he expressed himself or his fluency. This caused me a lot of stress, which again affected me because at certain beginning parts of the conversation, for example, I had to start the conversation myself or give my own ideas. (S4)

... I think we stress each other a lot among our friends. (S8)

It was a negative situation for me, I took it with my classmate last semester and he was someone I was very close to. That's why she calmed me down and I calmed her down. Now I was a bit nervous with people I didn't know. (S12)

The *exam questions* itself played a crucial role in shaping students' experiences. Not knowing the exam questions in advance was identified as a challenge since it prevented students from preparing thoroughly. Furthermore, difficulties in understanding questions due to remote reading or unclear formats as well as having new topics (though they were not new to the students since the topics were covered in the coursebooks and the class) in the questions added to the cognitive load during the exam, making it difficult for the examinees to respond effectively. Besides, the restricted time *duration* and the *exam room conditions* in which there were multiple examiners and video-recording existence, created additional pressure on students, making it harder for the students to demonstrate success in oral exams.

The most difficult thing for me was that they read the questions remotely, you know, first they read them once remotely and then they ask them in eye contact. (S2)

In the question styles asked, for example, this semester was a little better, but I had more difficulty understanding the questions with pictures in the first semester. Frankly, it was more difficult for me. Frankly, this semester, for example, was better for me because it was more like a dialogue. (S4)

Time. I lost a little bit of time because it doesn't give me time to think in the first question. What am I going to talk about? What will be the first sentence? This is a bit difficult. (S13)

In brief, from learners' standpoint, the struggles students cope with during speaking exams are diverse. While anxiety was cited as the most prominent challenge, negative peer effects such as unfamiliar exam partners and negative comments on the exam context, unsupportive and vague jury behavior, contextual issues related to the exam environment and vocabulary challenges including gaps in vocabulary knowledge and difficulties with complex vocabulary also made it hard for the students to perform well in the speaking exams.

4.2.2.3 Facilitators that affect students' speaking performance from students' point of view

Table 4.10 below shows the facilitators that had an impact on students' speaking performance from students' perspective. Several factors positively influenced students' performance during oral exams, as depicted in the same three themes: *Examinee-based*, *rater-based*, and *context-based* facilitators.

Table 4.10

The facilitators that affect students' speaking performances from students' point of view

Theme	Codes	Frequency	Description
Examinee-based	Readiness		Consistent practice helped students feel prepared and confident. Also, stress management, and the ability to focus on questions enabled students to perform better and stay composed.
	- Practice	10	
	- Well-understood lesson	2	
	Understanding question patterns	2	
	Knowing exam procedure	2	
	Self-perception		
	- Stress management	1	
	- Self-motivation	1	
	- Generating ideas	1	
Rater-based	Familiarity	5	Familiarity with the teachers, and their supportive behavior and clear guidance eased stress and enhance student performance during speaking exams.
	Behavior		
	- Encouraging & friendly attitude	5	
	- Asking questions understandably	1	
	- Clear pronunciation	1	
Context-based	- Giving prompts	1	Familiar topics and question patterns, knowing exam procedures, having picture-prompted questions and seeing them as well as having thinking time made
	Exam questions		
	- Familiar topics	3	
	- Use of visuals	1	
	Exam conditions		
	- Non-stressful environment	1	
	- Distributing questions	1	
	- Thinking time	1	
	Peer Effect		

Table 4.10 (continued)

The facilitators that affect students' speaking performances from students' point of view

-	Familiar partner	3	students feel more comfortable.
-	Not being influenced by the partner's performance & his/her question	1	

4.2.2.3.1 Examinee-based facilitators

Examinee-based facilitators involve readiness, understanding question patterns, knowing exam procedure and self-perception. Regarding readiness, which was cited many times, highlighted the positive effects of consistent practice in and out of the classroom as it reinforces students' confidence and increase readiness for the oral exams. Students also emphasized the significance of vocabulary practice, viz., to revise the previously learnt vocabulary items and build on what had already been learnt so that they would feel prepared for the oral tasks. Similarly, thorough understanding of the lesson content was deemed as another contributor to successful performance during speaking exam.

... and I studied a lot. (S6)

I think we can do more speaking activities. Both in class and among friends. (S8)

... vocabulary revision helps me to understand the lesson well and it helps me more when I do general revision or when I talk to any foreign friend or any English speaker in games. ... So practicing is the most important factor. I mean, if I practice a lot or if I talk to any of my friends and if I relax, it becomes simpler. (S9)

Studying for sure. Speaking regularly with my friends. For example, I think it improves when I practice speaking with my classmates a few days a week, even if not every day. (S15)

Students also emphasized the importance of understanding the question patterns and being aware of the exam procedure. Since these are the achievement tests, the question patterns were similar to the ones they already covered in their classes. Likewise, they

were already familiar with the exam procedure, which enabled them feel at ease during the exam.

... We know what we are going to face and we know what kind of questions we can expect because the questions are close to the units. This was the most important positive feature that put my mind at ease. (S2)

I think what makes my job the easiest is that we see all these questions and we are already doing these activities in the lesson because it is not the first time we have seen this. We have done these speaking activities in our previous tasks. (S5)

... what I've been saying from the very beginning is that there may be things that I've studied and that I'm actually familiar with seeing. (S11)

The final examinee-based facilitator is related to students' *self-perception*. Self-perception was regarded as another important enabler of better performance, and included managing students' own stress and maintaining self-motivation. With the help of this awareness, learners managed the exam process successfully and were able to stay composed during the exam. Besides, the capacity to generate ideas in an effective way also helped students achieve better outcomes.

I think I also manage stress well. (S6)

I always finish after 5 minutes in general and especially in speaking exams. After 2 minutes you will actually be relaxed. I say that if you are stressed right now, you cannot do it, you will be more upset because you cannot do it. So by making myself believe that it will be over in 5 minutes, I actually focus on that moment. (S11)

4.2.2.3.2 Rater-based facilitators

Examiner-related elements were noted as another key factor which facilitated oral performance. Encouraging and friendly attitudes and familiarity with the examiners were perceived as promoters since these factors created a less intimidating exam atmosphere, and in return, lessened students' anxiety. Teachers' giving prompts, understandable question phrasing and clear pronunciation also enhanced students' ability to perform well, as they assisted students' comprehension of the questions and instructions thoroughly.

You know, they say they good or something, their reactions or something like that, for example, when they relax a little bit, it makes it easier (for us) to construct sentences... (S7)

Yes, when there were definitely professors I knew around, I felt more comfortable. I mean, I think I express myself more easily because I feel less stressed. (S14)

... you can be better motivated when there are familiar teachers. (S10)

First of all, my teachers. It was easy for me because the questions were very clear, the words were pronounced in an understandable way. Understanding is the most important factor. (S1)

4.2.2.3.3 Context-based facilitators

Context-based facilitators include *exam questions*, *exam conditions* and *peer effect*. Students mentioned the characteristics of the *exam questions* since they could ease students' stress and enhance their performance. To be more specific, familiar topics in the questions were regarded as key factors that helped students feel more comfortable during the exam. In addition, the inclusion of visuals in questions were also stated as positive elements. Second, providing students with non-stressful *exam conditions*, handing the examinees the questions so that they could read the questions themselves, and allowing them time to think before responding were also underscored as positive practices that contributed to better speaking performance.

I knew what I was going to face, I knew how the exam was going on because similar videos are shown in the classroom beforehand. Anyway, this is a very important factor, we know what we will encounter, we know what kind of questions may come, because the questions that are close to the units are already coming. (S2)

I think what makes my job the easiest is that we see all these questions and we are already doing these activities in the lesson because it is not the first time we have seen this. We have done these speaking activities in our previous tasks. We are constantly talking to our partners with similar questions or during the lesson. I think these speaking activities make our job easier. (S5)

One of the positive things is that they show the questions again, which is very comforting. At least we have the opportunity to see this question. The fact that they give us the opportunity to think afterwards and give us the opportunity to analyze the question is actually a very positive factor. (S2)

Peer effect also played a pivotal role in students oral exam performance. Particularly, the presence of a familiar partner minimized negative elements and fostered a collaborative environment during the exam. Furthermore, not being influenced by the partner's performance and the question that corresponded to him/her minimized unnecessary pressure and helped them focus on their own responses.

I think the person I am matched with ... I mean, if I know the person I am matched with and we are at a close level, then I really feel much more comfortable because, you know, in the discussion part, for example, if I cannot start, he can start, he can advance. (S4)

Since two of you are taking the exam, you should not be affected by the questions of your other friend. Some of my friends could not do it, but I thought it was easy when I heard his questions, but my question was more difficult for me, but I still answered it. (S1)

To summarize, the findings gathered from students' interviews emphasize the major facilitators that contribute to a supportive environment for students during speaking exams. The most prominent of all are readiness for the exam with constant practice, taking the exam with familiar raters' and their encouraging and friendly demeanor. Similarly, familiar topics in the exam tasks along with familiar exam partners could help students feel at ease and excel in oral exams (see Appendix I for the overall findings of both instructors and student interviews).

4.3 Video-stimulated Recall Protocols

Beside semi-structured interviews, video-stimulated recall (VSR) interviews were conducted with both EFL instructors and students to gain deeper insights into their perspectives on the speaking exam process. Students reviewed the video recordings of their own. The instructors viewed their students' midterm speaking exam performances, and all participants reflected on their experiences. By revisiting these moments, the interviews provided a unique opportunity to explore the cognitive and emotional aspects

of the exam from both teachers' and learners' standpoints. For instructors, it shed light on their assessment practices and decision-making processes as well as the factors that influenced the examinees' performance. For students, it revealed their thoughts, challenges, facilitators and strategies during the exam.

4.3.1 Video-stimulated recall interviews with the instructors

This section presents the findings obtained from the VSR interviews conducted with the instructors. To gain a more comprehensive understanding of the instructors' perspectives in relation to the research focus, this section is organized under two main headings: factors that affected students' performance during the exam and what instructors would do differently if they were to re-administer the exam.

The summary of these findings was displayed in Tables 4.11 and 4.12. The excerpts from the instructors VSR interviews are labeled as VSR-I1, VSR-I2, and so forth. The interviews were conducted in the native language of the participants; thus, the translations were done by the researcher.

4.3.1.1 Factors that affected students' speaking performance

Table 4.11 below explains the key elements that had an impact on students' performance during midterm speaking exams. Similar to the semi-structured interviews, three main themes unfolded: *Examinee-based*, *Rater-based* and *Context-based factors*.

Table 4.11

Factors that affected students' speaking performances from instructors' point of view

Theme	Codes	Positive / Negative	Frequency	Description
Examinee-based	Readiness			A strong vocabulary and grammar as well as completing the assigned task positively impact performance, while a limited range can hinder expression.
	- Task completion	+	3	
	- In-class performance	+	1	
	- Memorizing speaking task questions done in class	+	1	
	Linguistic factors			
	- Grammar complexity	+	2	
	- Vocabulary range	+	2	
	Affective factors			
	- Anxiety	-	2	
	- Mood	+/-	1	

Table 4.11 (continued)*Factors that affected students' speaking performances from instructors' point of view*

Rater-based	Familiarity	+	3	Being familiar and a friendly demeanor or a strict and monotonous speaking from the examiner help or intimidate students while performing.
	Behavior			
	- Attitude (smiling, strict, monotone speech)	+/-	1	
	- Tone of voice	+/-	1	
	- Being a high/low grader	+/-	1	
	- Understanding student abilities	+	1	
Context-based	Criteria	+/-	2	Being recorded, noise from outside the exam room during the exam and feedback from earlier examinees can distract and reduce focus.
	Exam conditions			
	- Video recording	-	1	
	- External noises (distractors)	-	1	
	Peer Effect			A strong and a familiar or supportive pair can motivate the student.
	- Pairing dynamics	+/-	2	
	- Familiar partner	+	2	
	- Comments from previous students	-	1	

4.3.1.1.1 Examinee-based factors

Instructors emphasized the importance of *examinee-based factors*, particularly *readiness*, *linguistic factors* and *affective factors* as a significant determinant of success in oral exams. In terms of *readiness*, instructors highlighted the importance of task completion, students' in-class performance and doing memorization exercises, all of which helped students be ready for the exam. Namely, when students could fulfill the requirements of the exam tasks properly, they were highly likely to achieve a successful outcome. Likewise, when students actively participated in in-class activities and prepared for the exam tasks through consistent practice and memorization exercises, given that it was an achievement test aligned with the course content, successful results were inevitable. In addition, linguistic factors such as vocabulary range and grammar complexity are positively correlated with performance while a limited range of these linguistic features hindered expression, and hence, successful performance. Finally, examinees readiness was determined by the affective factors including anxiety and mood. Regarding affective factors, anxiety was seen detrimental, as it hindered focus and the ability to express ideas well enough. In a similar vein, a student's mood on the exam day could influence their

confidence, willingness to engage in the task and as a result, their performance. The students who effectively managed their day and maintained a positive mood demonstrated improved performance in the exam, whereas the students with a low mood exhibited a decline in performance.

... I mean, the student uses vocabulary, use what we use in class, or do they all say good, fine, bad all the time, or do they use more complicated structures? If not, I mean, they used everything in grammar without any mistakes, but could they be better? (VSR-I3)

... not only did they fulfill the task, but they also used various complicated structures on top of it. (VSR-I3)

... whether their vocabulary is good or bad, whether they study grammatical structures or not... (VSR-I5)

I think there is a tendency in our children to memorize questions, for example, when they hear the questions that we do in speaking activities a little bit differently, for example, or they have studied a very similar question, they have memorized it. He gives the answer to it a little bit. They come and explain the question they prepared for, the answer is not exactly to the point, but it is somehow relevant, for example. (VSR-I6)

Well, of course the anxiety must be affecting (them). (VSR-I7)

Of course, I think the mood of the day, I mean it can be very effective. Maybe he got bad news or he got a happy news, that can affect it. (VSR-I4)

4.3.1.1.2 Rater-based factors

Rater-based factors were the second emerging theme, consisting of *familiarity* with the jury members and their *behavior* during the exam. Familiarity with the examiner was perceived as a positive effect since it relieved the learners during their performance. Similarly, examiners' positive behavior such as demonstration of an understanding of students' capabilities affected students' performance positively. In contrast, the examiner's attitude and tone of voice had both positive and negative effects on student performance. More precisely, the examiner's friendly demeanor positively impacts

performance, whereas strict or monotonous behavior and the tone of voice intimidated students, which affected their confidence and ability to speak fluently.

If his own teacher is entering, maybe he is happy at first. If not, there is a possibility that they are a little more serious, a little more stressed. It could be the other way around, when I see that mine (my students) seem to be happy. (VSR-I9)

Knowing what their level is. They had that kind of discussion part, for example, but we think that this student is not capable of defending this in this way, and we actually grade the student accordingly, although it is more of a discussion, as you can see, no student in level D can actually make a discussion in general. (VSR-I8)

I also heard that when we entered, the teacher was not friendly at all and very strict. I mean, he directly asked the question in such a monotonous tone and we were very tense. I heard the opposite of what I just said. This can also be an effective factor. (VSR-I4)

Even sometimes when my partner is not a man but a woman, I say, you should greet them with a tone of voice, at least so that they are not startled. (VSR-I1)

4.3.1.1.3 Context-based factors

Final factors that affected students' performance were categorized under the theme of *context-based factors*. These factors involve criteria, exam context and the peer effect. Firstly, *criteria* were seen as both a positive and negative influence. Meeting these criteria was crucial for students to achieve higher scores, as it aligned their performance with the examiners' expectations. Therefore, the effective use of criteria during exams was critical from the instructors' view. Instructors believed that it was important to stick to the criteria while assessing students' performance in order to avoid subjectivity, and maintain the standards among all examiners, and thus, be fair.

Criteria. I mean, I looked at the criteria, I looked at grammar, I looked at word choice, I looked at whether he could form sentences, I looked at whether he could express himself, I looked at whether he could develop an answer to that question, I looked at his pronunciation, I looked at them, I generally look at the things that are in the criteria. (VSR-I7)

Second, *exam conditions* were seen as another significant effect. Instructors, namely, pointed out the negative elements in the environment in which the speaking exams took place. One of these negative factors mentioned was video recording, coupled with external noises from outside the exam room created distractions and affected concentration. The presence of such disturbances made it challenging for students to perform better in the exams.

It (Recording) may excite them, but in the end, it is normal for something like that to happen at a time when they want to object. Actually, they are aware of it because they have been informed in advance. I mean, it's also because this is how it should be, but after all, there is a record, they may get excited. (VSR-I1)

... In other words, we are trying to protect them not only in the airplane but also against external noises, in other words, for a healthier environment. (VSR-I2)

Finally, *peer effect*, specifically the dynamics of pairing, had both positive and negative influences on students' performance. Instructors viewed that a familiar or supportive partner motivated and encouraged students; conversely, a less suitable pairing negatively affected performance. Also, prior to the exam, comments from previous examinees created tension among the students.

From the student's point of view, as we mentioned, the pair they enter together can be an effective factor. (VSR-I4)

One pair is good and the other is not. So it doesn't affect it very much. It may push forward a little bit in a positive direction, but it doesn't seem to change the result too much. (VSR-I1)

By the way, it may also affect who they take the test with, that is, it also affects if they enter with a classmate they like. There may be someone they don't like very much in the class. There may be someone they are ashamed of, I think that also affects their performance, that is, their performance can also change in front of them. (VSR-I6)

Also, when these students leave (the exam room), they talk to each other and I'm not sure if this is something that relaxes them, but what came out, who

was there, how it went, you know, maybe if they get positive feedback from the student who left, the tension in that student may disappear or decrease a little bit. Therefore, I think that there is a dialog between the first and the fourth student after they leave, and I think this can also be a factor. (VSR-I4)

In sum, various factors influenced students' speaking performances from instructors' point of view, ranging from individual factors to factors stemming from examiners and exam context. Individual factors comprise students' readiness, their linguistic competencies and affective factors. Examiner-based factors mainly encompass familiarity with the examinees and their behavior during the exam. Context-based factors are related to the exam criteria, exam setting and peer effects. These insights highlighted the areas in the exam design and preparation which influenced students' success either positively or negatively.

4.3.1.2 Suggestions for re-administering the exam

In VSR interviews, the instructors were also asked to provide suggestions for potential changes if they were to administer the exam again. Table 4.12 presents these recommendations to improve the administration and the structure of speaking exams so that the students can reflect their real oral performance. The suggestions were categorized into two main themes: *Context-based* and *Rater-based suggestions*. While context-based suggestions basically encompass exam questions and the administration, rater-based suggestions involve familiarity with the testers.

Table 4.12

Suggested changes for re-administering the exam

Theme	Codes	Frequency	Description
Context-based	Exam questions		Modifying question roots or instructions to make them clearer and easier to follow, as well as eliminating the discussion part, might be better. and keeping pairs or solo performance to avoid dependency or unfair dynamics in pair tasks.
	- Revision of the wording of the questions & instructions	4	
	- Removing the discussion section	2	
	Exam Administration		
	- Individual testing instead of pair/group testing	1	
	- Avoiding mouth masks while administering the exam	1	

Table 4.12 (continued)*Suggested changes for re-administering the exam*

	- Providing help to students struggling with questions	1	
Rater-based	Familiarity with the instructor	1	Since it boosts students' confidence, some teachers prefer to administer the exam to their own classes
No change		2	

4.3.1.2.1 Context-based suggestions

In terms of *context-based suggestions*, instructors suggested some adjustments in the structure and the clarity of the *exam questions* and the *exam administration*. To be more specific, some instructors believed the current wording and the instructions of the questions were restrictive. Namely, they made it difficult for the students to generate ideas, and thus, fulfill the task requirements adequately. Thus, they explicitly suggested to revise the wording of the questions and the instructions. Furthermore, two instructors emphasized the need for removing the discussion section in the exam, where students were expected to discuss, share opinions, use certain strategies such as turn-taking and agreeing-disagreeing. According to the examiners, this section might unnecessarily complicate the test for students. Hence, students would reveal their capabilities better thanks to the adjustments related to the exam content.

Sometimes the questions, you know, sometimes I put myself in the shoes of the students. If they ask me, what can I say? Sometimes in the class, I can't think of the Turkish version of this, so I should say it in English. So, some questions can be absurd. As soon as they say yes, they have given the answer to the question, the question may not go any further. (VSR-I1)

For example, I said there shouldn't have been a discussion part. That would be the part I would want to change. ... that is, a student should answer two questions and finish the exam. Actually, I think this is enough for D (level learners). (VSR-I8)

Another context-based suggestion made by the instructors was related to exam administration. Instructors believed that certain adaptations about the current implementation which restricts students' demonstrating their capacities are necessary.

One suggestion involves individual testing. According to an instructor, shifting from group testing to individual testing when a partner is unavailable might reduce dependency or potential unfair dynamics in pair tasks. Additionally, one instructor felt uncomfortable during the exam since he had to wear a mask due to his illness as it might hinder the communication during the exam. Also, the desire to offer help to students who struggle with questions by not repeating the question three times but being able to paraphrase it was articulated as a valuable suggestion. They thought repeating the question with the same words was useless and anxiety-provoking since it did not lead to comprehension. By the same token, although teachers wanted to rely on evaluation criteria for assessment, a more elaborate and user-friendly set of criteria would be more effective.

It (Implementing the exam individually) can be done, it doesn't even have to be in pairs. Student can come alone. (VSR-I8)

I was sick that day and I was wearing a mask. I think that could be a factor. I mean, I could have changed that, but of course I can't help it. (VSR-I4)

What needs to be changed is the criteria, the evaluation criteria. Whatever I do, I cannot give a low grade to the student. Either I give a very low grade to all of them, one, one, one, one... Either because they do not fulfill the assignment or because they already fulfill it, I give them four, four, four, four, four, four, I give them all a high grade of 80, that is, the student gets a score 20 points higher than the passing grade. (VSR-I3)

I would also like to be involved because there may be a word that the child does not know. It may be necessary to open it, we can explain it somehow because they are experiencing an exam anxiety inevitably. (VSR-I5)

4.3.1.2.2 Rater-based suggestions

Beside context-related suggestions, one instructor recommended implementing oral exams with the familiar examiners. He believes that ensuring familiarity with the teacher might boost students' confidence, ultimately leading to improved performance. Therefore, it is suggested that teachers administer speaking exams to their own students to create a more comfortable exam environment and mitigate potential negativities associated with unfamiliar testers.

... When students who know them come, they (students) can feel a little happier, a little more confident. (VSR-I1)

It should also be noted that although these issues were emerged after the VSR interviews with the instructors, two of them stated they would change nothing. One also explained that these achievement tests have been conducting for many years, and necessary changes and adjustments are already being made as needed.

... no, because we already have an established system and we constantly receive feedback and update it when there is a problem. (VSR-I4)

... The questions were asked in a good way. For example, things were not misleading. The pictures were not misleading, they were also beautifully and carefully selected, in that sense, there was nothing I could change, nothing I would want to change. There was nothing that came to my mind that I wish it was like this. (VSR-I2)

All in all, Table 4.12 reflects practical suggestions to enhance the effectiveness of speaking exams. The instructors' recommendations include modifications to the exam content such as revising questions and removing restrictive elements, to the exam administration including familiar examiners and clear communication, and a more flexible evaluation approach in which teachers can provide support to struggling students so that students can demonstrate better performance and achieve higher success.

4.3.2 Video-stimulated recall interviews with the students

The VSR interviews were also carried out with three students across five language levels, totaling 15 students, who took their midterm speaking exam two weeks before the interviews. The data gathered from these interviews took approximately 200 minutes, around 15 minutes per student, and were categorized under two themes: factors that affected their performance and what they would do differently if they were to retake the exam. Before watching the video recordings, students were asked whether they had found their speaking exam score fair enough. One student stated that her grade was not fair enough at the beginning of the interview, but changed her mind after watching her own performance, so 12 out of 15 students reported that they found his/her grade fair. Below are some reasons why they found it fair and unfair:

I mean, I messed up a bit on the first question, which was bad. I spoke moderately in 2 but I think I spoke in 3. I got 50, which is right in the middle, and I think it is an ideal number. (VSR-S9)

I mean, not too much, not too little, I mean, it could have been a little more, maybe in terms of time, I tried to speak fluently. Maybe the effort could have been given a little more. (VSR-S7)

4.3.2.1 Factors that affected students' speaking performance

Apart from fairness, factors that affected students' speaking performances during the exam were also asked, and various elements influencing students' performance during exams emerged. These positive and negative factors are grouped into three main themes, which are *examinee-based*, *rater-based* and *context-based factors*.

Table 4.13

Factors that affected students' speaking performances from students' point of view

Theme	Codes	Positive / Negative	Frequency	Description
Examinee-based	Affective factors			Affective factors such as anxiety and nervousness along with self-perception and linguistic factors had a significant impact on students' ability to perform effectively.
	- Anxiety	-	6	
	- Fear of misunderstanding the question	-	2	
	- Peer influence concerns	-	1	
	- Feeling shy	-	1	
	- Feeling unprepared for questions	-	1	
	- Fear of low grades	-	1	
	Self-perception			
	- Answering questions successfully	+	2	
	- Being more fluent	+	1	
	- Complicating simple situations	-	1	
	- Fear of poor communication	-	1	
	- Fear of making mistakes	-	1	
	Linguistic factors			
	- Lack of vocabulary	-	2	
	- Difficulty forming correct sentences under pressure	-	1	
	Attention deficit	-	1	
	Arrival time	+/-	1	

Table 4.13 (continued)*Factors that affected students' speaking performances from students' point of view*

Rater-based	Familiarity	+/-	4	Familiar and supportive teachers eased student anxiety, while unfamiliar or unapproachable teachers created discomfort and reduced confidence.
	Supportive behavior	+	2	
Context-based	Peer effect			A calm environment improved comfort, while noise and chaos during the exam distracted and stressed students. Using visuals in questions promote the understanding, though.
	- (Un)Familiar pair	+/-	4	
	- Examining with three people	-	2	
	- Late partner	-	1	
	Exam Environment			
	- Relaxed classroom environment	+	1	
	- Noisy environment (airplane noise)	-	1	
	- Crowded, chaotic pre-exam atmosphere	-	1	
	Exam conditions			
	- Lack of thinking time	-	1	
	- Time of the exam	-	1	
	- Participating after peers	+	2	
	- Distributing questions	+	1	
	- Taking the exam after his/her pair(s)	-	1	
	Exam questions			
	- Picture-prompted questions	+	1	

4.3.2.1.1 Examinee-based factors

Examinee-based factors encompass affective factors, self-perception, linguistic difficulties and some other personal factors including attention deficit and arrival time. To start with, *affective factors*, mainly anxiety and nervousness, were the most frequently reported negative factors, mentioned by the students. These emotions hindered students' focus and natural expression. In addition, specific concerns such as fear of misunderstanding the question, peer influence, shyness, feeling unprepared and the fear of receiving low grades were also viewed as negative impacts that created high pressure.

I'm going to say it again, but nervousness. I can see the tension on myself. I mean, I never thought I was so nervous during the exam. (VSR-S5)

I probably entered (the exam room) after them (exam partners), and in fact, when I watched this video, I didn't enter that late, so the signatures are just being signed. But when the people at the door told me that your people (exam partners) had already entered, they had already started, you know, I stressed that even more. (VSR-S12)

In fact, I get even more stressed when I talk to a friend because I think I am influencing her at the same time. And the anxiety that we won't communicate well enough at work and that we will get a lower grade also affects my stress. (VSR-S14)

On the other hand, students' *self-perception* influenced their performance either positively or negatively. To illustrate, when students believed that they answered the exam questions successfully and were fluent enough, they minimized self-doubt and perceived themselves successful, and thus, felt relaxed throughout the exam. On the other hand, complicating simple situations, fear of failure and making mistakes were also cited as negative factors, which hindered successful performance.

One of the factors that affect me negatively is that I misunderstood the first question and then stress (I felt stressful) when I answered the rest of the questions. (VSR-S9)

Another examinee-related factor was *linguistic factors* which were also perceived to affect real-time performance adversely. In particular, lack of vocabulary and having difficulty in forming correct sentences under pressure were seen as significant barriers to effective communication. Also, some personal issues impacted their performance. For example, attention deficits and arrival time were deemed as further complicating elements since they affected students' ability to focus and respond accurately.

First of all, the lack of vocabulary definitely affected me a lot, and apart from that, the question asked was future tense, but I spoke in present tense and present tense, which was a big mistake... I would have been able to make more sentences if I had studied vocabulary. (VSR-S15)

A factor that affected me in a different way during the exam was my lack of attention, which I have already told you before. That's why not understanding the question correctly made me nervous. (VSR-S2)

4.3.2.1.2 Rater-based factors

In addition to examinee-based factors, *rater-based factors* were mentioned as influences on students' speaking exam performance, as well. These include familiarity and supportive behavior. That is, students stated that acquainted and supportive examiners eased their negative emotions, while unfamiliar teachers created discomfort and reduced confidence.

... the teachers sometimes did a lot of things, you know, they looked like this. You know, okay, they have to do that. When I looked at them, I felt something like this, I thought I must be doing something wrong and asked the question twice more, for example, I wondered if I misunderstood, if I was talking about something else. (VSR-S7)

Of course, having familiar teachers, my friend being my classmate. It affected me positively because going with familiar teachers already knows me, my teacher knows how I talk in class. It makes you feel like you don't need to be nervous. (VSR-S6)

...Having teachers I know. It had a positive effect. ... I mean, I felt more secure, for example, ... You know, at least seeing someone we know in front of us makes us feel more secure. (VSR-S11)

For example, you know one of your teachers, but not the second one. That's why I felt a bit embarrassed. (VSR-S13)

4.3.2.1.3 Context-based factors

Context-based factors were seen as other influential elements in students' speaking exam performance, and involve peer effects and exam environment. Regarding *peer effects*, four students mentioned the positive effect of familiar peers since it reduced their anxiety, whereas unknown pairs were viewed as stress-inducing. On the other hand, taking the exam in a group of three instead of in pairs, and late-arriving partners were reported to

have adverse effects on students' performance, likely because these factors increased anxiety, disrupted concentration, and created an imbalance in interaction dynamics.

I think being (taking the exam) with someone I don't know affected me. If it was someone I knew, I could talk a little more comfortably, I mean, it's someone I know, I don't know, he knows how I would talk, I know how he would talk. I think there could have been a more comfortable conversation between us. (VSR-S3)

...um, at first, the friend I matched with came a little late, not too late, but later than I expected. I was already nervous there. (VSR-S4)

... and it was even more stressful because there were three of us. Two of us could talk by looking at each other, but when there are three people, when you say something, two of your friends next to you and two lecturers look at you. That's why it was even more stressful. (VSR-S13)

In the same way as taking the exam with my classmate, sometimes, you know, we can match from other classes. That puts you in a bit of trouble because especially in the discussion section, will he help or what will he do, for example, here we are helping (name). At the point where she is blocked, I enter, she passes it to me, something happens, but when I enter with someone I don't know, there is also the stress of whether it will happen or what we will do. (VSR-S6)

Exam environment also played an important role in students' ability to focus and perform. A relaxed classroom environment was cited as promoting comfort and confidence. Yet, various factors such as noise (e.g., airplane noise), a crowded and chaotic pre-exam atmosphere, insufficient time to think about the questions and the time of the exam were seen as negative influencers.

...the classroom environment's being a little more relaxed, that is, not too suffocating (VSR-S1)

Afterwards, when we went up to the classroom, these windows were open and at first we heard the sounds of airplanes and then we closed the window. (VSR-S4)

Similarly, *exam conditions* also impacted performance. To start with, students' felt not having enough time to think about the question before they answered. Also, the time they took the exam affected their performance. For example, some students did not prefer to take the exam in the early morning sessions as it negatively affected their overall test achievement. On the contrary, some students mentioned the effects of answering the exam question after his/her partner(s). The questions given to the students was also yielded as a positive outcome for the examinees.

It is also important not to extend the time to think too long, but it can be extended a little longer, or the student can start when he or she feels ready to start. (VSR-S2)

That's why I was nervous about not understanding the question correctly, but they also share the question with you so that you can read it again. So, it didn't affect me negatively. (VSR-S2)

When your friends come in and out, we talk to someone, you know, what did they ask, how was it, you know, while talking, I relax a little more and I get used to that exam atmosphere a little more. (VSR-S11)

Likewise, *exam questions* especially regarding the design and delivery of the exam questions were seen as another influential factor. Specifically, when the questions were prompted by related pictures and were handed to the students, they found it more helpful in terms of comprehension.

It was better to have pictures, we can do something by looking at them to explain. For example, normally, if that question had been put, I would have just said 1 or 2 factors. It was over, I wouldn't have said it after that. It gave more ideas, yes. (VSR-S13)

To sum up, Table 4.13 below identifies a range of factors which were at play while students were performing. Not only familiar exam partners but also familiar instructors were believed to ease the process for the test takers. In addition to familiarity, supportive and helpful teachers were also seen as a facilitator. Also, exam questions as well as relaxed exam environment were perceived as fostering elements during the exam, whereas negative feelings such as anxiety and self-perceptions as well as language barriers were mentioned as challenges.

4.3.2.2 What students would do if they were to retake the exam

Students were also asked what they would do differently when they were given another opportunity to re-take the exam. The following table (Table 4.14) describes the strategies students would use to perform better. Although a variety of strategies appeared to be available, they could be broadly categorized under the following themes: *examinee-based* and *contextual strategies*.

Table 4.14

What students would do if they were to retake the exam

Theme	Codes	Frequency	Description
Examinee-based strategies	Practice (general)	13	Students
	Expanding & improving vocabulary use	5	highlighted the
	Reducing stress and anxiety	4	need to improve
	Using thinking time effectively	4	fluency, accuracy,
	Avoiding haste	3	vocabulary,
	Using professional and quality sentences	3	grammar, and
	Speaking fluently	3	pronunciation
	Practicing pronunciation	2	through
	Recording oneself to improve fluency & pronunciation	1	preparation and
	Avoiding the mindset of failure	1	consistent practice.
	Avoiding eye contact with teachers to stay calm	1	Also, they
	Trying to make amends for wrongdoing	1	suggested
	Showing more of what s/he knows	1	strategies for
	Thinking simple	1	managing their
	Interacting with peers more effectively (in discussions)	1	own responses and
	Using better pronunciation	1	behaviors to reduce
	Expanding my speech	1	pressure and
	Using grammar more carefully	1	improve outcomes.
	Practicing fluency	1	
	Getting to know the pair before the exam	1	
Context-based strategies	Starting as the second student in pairs	1	

The most prominent theme was examinee-based strategies, which primarily consisted of individual approaches to test-taking. First of all, a majority of the test-takers emphasized the significance of being prepared for the exam. Especially engaging in both in-class and out-of-class exercises, particularly those focused on pronunciation and fluency development, can contribute to the improvement of both fluency and accuracy.

Now it's like this, these questions are probably, no most likely, they are questions from the topics of the book, of course, I listened to it in class and I

did speaking activities with my friends and teachers, but I could go back and look at these questions again before the speaking exam because frankly I don't study for speaking exams. (VSR-S5)

I would try to practice a little bit more on the questions that might be asked beforehand, about speaking. ... but I think I could have done something outside the classroom, maybe with my friends, maybe by myself, I could have practiced. (VSR-S15)

Use of strategies was also considered valuable for the students. In order to be able to manage their own responses, some students suggested some plan of actions such as reducing stress and anxiety, using thinking time effectively, and avoiding haste at the time of speaking. Some individual strategies were also suggested by the learners including recording themselves to improve their fluency and pronunciation, showing more of what s/he knows, and thinking simple. Recording oneself to assess and refine fluency and pronunciation was suggested as an additional strategy to perform better. In addition, avoiding the mindset of failure and eye contact with the test administrators were also suggested by the learners to stay calm during the test. That is, students emphasized the necessity of dedicating time to improving fluency and accuracy in pronunciation and language use. Other strategies involve peer dynamics such as having a more effective interaction in the discussion task and getting to know them before the exam. Namely, students valued getting to know their pair before the exam to build rapport and enhance collaboration during the exam.

... and I would also try to calm my nervousness a little more because I repeat myself a lot in what I say, I mean, I say the same thing after two sentences. (VSR-S4)

First, I would not stress. (VSR-S11)

I'd like to go over it (my anxiety). (VSR-S15)

I'm not supposed to fall into the trap of ..., I shouldn't get into the psychology that I can't do it right now. I would do it differently, I would be a little more comfortable because you need to give that comfort to the other side. It is not only something that comes from the teacher. If you are also very tense, it affects them too, because dialog is such a thing, both sides need to stay within

the flow of the dialog in this process. If I am comfortable, the other side will be comfortable... I would like to be the second student in the next one. (VSR-S2)

When the questions are asked, we generally have 30 seconds of thinking time. I usually don't use the 30-second period. I mean, when I have a question in my mind, it determines something directly in my mind and I start talking directly, but if I use 30 seconds, maybe I can plan things more clearly, I can sort things in my mind and sentences better, I don't fall into repetition. (VSR-S4)

I would think more deeply about the questions and not rush to answer them. I hope by thinking more. Without being in a hurry. (VSR-S1)

By opening up the subject more, that is, by increasing and developing it in terms of content and developing good sentences, for example, in this exam I tried to speak (use) more words, but in the next one, I think it is better to make more quality, at least more fluent and proper sentences. (VSR-S2)

And at the end, it would have been better if there was a little more communication in the conversation. (VSR-S3)

I know more, in fact, I was supposed to show what I know, but you know, for example, I showed 75-80% (of what I know), if I had shown 100%, I would have gotten 100 (points). (VSR-S6)

I would meet those people before the day I was going to take the exam, I would go up to them and say hello or they were already in the next class. I could stop by the classes and ask, or there are people I know in the next class, you know, I could go to them and meet them, I could do something, that would reduce my stress more. It would minimize it (the stress). (VSR-S12)

I used to think a little bit simpler, you know, ... I would put them (the sentences) all in the same order, one after the other, so now, for example, as I said, for example, I played basketball, I wouldn't think that I was penalized. For example, I was playing. It was very good, I was with my friends, etc. ... I probably wouldn't look at the teachers, I would probably look somewhere else, that's what influenced me a little bit. (VSR-S7)

Finally, students also highlighted some strategies related to linguistic development. As cited by one-third of the participants, expanding and improving vocabulary use was noted as a highly important strategy for better performance. Using professional and quality sentences along with speaking fluently were also cited three times, which is followed by using better pronunciation. In addition, trying to expand their speech and use grammatical structures more carefully were cited by students for producing more thoughtful and accurate responses.

I think I would have lengthened my speech a little bit more because I think I spoke too short. I think I got a lot of points for my short speech, and I was very stuck in the intervals. If I hadn't said that, I would have spoken a little more fluently, for example, I would have said "I mean" or something like that. I think maybe I could have pronounced it a little better there. And at the end, it would have been better if we had communicated a little more in the conversation with the other person. (VSR-S3)

Briefly, students view readiness, use of strategies and language skills development as pivotal for improving their speaking performance in exams. Practical strategies such as practicing language skills for both fluency and accuracy, building confidence by reducing stress, managing time effectively, and fostering better peer interactions are considered as central strategies to achieve better outcomes (see Appendix J for the overall findings of both instructors and student VSR interviews).

5 CONCLUSION, DISCUSSION and SUGGESTIONS

5.1 Conclusion

The purpose of this study was to identify the factors that influence EFL students' speaking performance during achievement tests, drawing insights from both instructors and students. Utilizing three primary data collection methods, which are questionnaires, semi-structured interviews, and video-stimulated recall (VSR) interviews, the research presents a nuanced understanding of these factors, offering a comprehensive view from both perspectives. The findings reveal both overlapping and divergent points, highlighting shared insights as well as differing emphases on specific factors impacting speaking performance.

The instructor questionnaire, which included responses from 98 EFL teachers, highlighted *Vocabulary Range* as the most significant influential factor with a mean score of 2.81, indicating that a robust vocabulary is crucial for effective communication from the instructors' perspective. Other notable factors included *Motivation to Speak* and *Vocabulary Appropriacy* with mean scores of 2.78 and 2.76, respectively. Conversely, factors deemed less influential encompassed the context of the exam such as *Rooms physical conditions* ($M=1.52$), *Exam slot time* ($M=1.44$) and *Order of Performance* ($M=1.20$). In contrast, the student questionnaire, gathered from 210 students, echoed different ideas. Namely, factors such as the *Number of Tasks* ($M=2.35$) and *Task Type* ($M=2.34$) were perceived as critical to their performance, whereas *Exam slot time* ($M=1.77$), *Fear of not being understood by the partners* ($M=1.76$) and *Partner(s) in the exam* ($M=1.75$) were deemed as less critical elements. In brief, the data suggests an understanding that while the instructors deemed the importance of vocabulary as a crucial factor in the speaking assessment context, EFL learners highlighted the task clarity as the most influential elements. Both groups indicated that external conditions such as *Fear of not being understood* and *exam context* factors were less critical in comparison to internal influences over vocabulary and motivation, reinforcing a consensus on the foundational role of linguistic competencies in effective speaking.

Moving beyond quantitative data analysis, semi-structured interviews with 22 instructors and 15 students as well as VSR interviews with nine instructors and 15 students enriched this understanding by delving into the qualitative aspects of these factors. These in-depth interviews provided richer qualitative insights into the dynamics of EFL learners'

speaking performance. Particularly, semi-structured interviews unveiled the categorization of factors into three main themes: examinee-based, rater-based, and context-based factors. Examinee-based factors primarily reflected internal influences such as anxiety and linguistic proficiency, which instructors identified as crucial for successful performance during oral exams. Rater-based factors were noted as significant, particularly how the demeanor and familiarity of examiners could foster or diminish students' confidence and willingness to perform. Furthermore, context-based factors, including the exam environment, types of tasks, and clarity of evaluation criteria, played a crucial role in shaping students' experiences and performance outcomes. Students' interviews also largely supported these themes, particularly emphasizing readiness, anxiety levels, and the behavior of raters as pivotal to their speaking exam success. Many students highlighted the comfort accompanied by familiar peers and supportive examiners, which lowered their anxiety and allowed them to display their abilities more efficiently.

Further exploration through video-stimulated recall (VSR) protocols with both instructors and students provided additional layers of insight. Instructors reflected on various examinee-based factors, such as readiness and affective states, as critical determinants of success in oral exams. They pointed out the importance of task fulfillment and in-class participation for effective preparation. Rater-based factors, including examiner behavior and familiarity, were also underscored as influential, with instructors noting that supportive and positive interactions could greatly benefit student performance. Contextual factors remained crucial, with instructors expressing concerns about the clarity of exam criteria and the overall testing environment, stressing the need for clarity in exam questions and potential adjustments to enhance student performance. Similarly, student responses in the VSR interviews revealed the complexity of their experiences during speaking exams. They identified anxiety as a primary hurdle, along with self-perception and linguistic challenges, which adversely influenced their ability to perform well. Nonetheless, learners also indicated the positive effects of familiarity with their peers and the examiners, alongside the importance of a relaxed exam environment, which could boost their confidence and diminish stress. Regarding potential improvements, students suggested strategies aimed at enhancing their readiness, such as more focused

practice on pronunciation and fluency, as well as building rapport with peers prior to examinations to create a collaborative atmosphere.

Despite these overlapping insights, the data sources demonstrated some differences in emphasis, as well. The quantitative data suggested that instructors pointed out vocabulary-related factors, whereas students placed particular weight on their personal experiences of anxiety and readiness as more influential in their performance. This contrasted with instructors' broader focus on linguistic competencies, suggesting a possible misalignment between perceptions of vocabulary applicability in an exam context versus actual student experience under pressure. Furthermore, while both instructors and students acknowledged the importance of the exam environment, the students' insights into specific contextual factors - such as pre-exam peer dynamics and the implications of timing - emphasized the stress-inducing nature of unfamiliar partners and exam settings in a way that the instructors did not fully recognize.

Consequently, the integration of questionnaire data, semi-structured interviews, and VSR reflections creates a comprehensive picture of the factors affecting speaking performance among EFL learners. The findings show the complex nature of speaking assessments and the interplay among various factors that influence EFL learners' speaking performance. The interplay between internal factors such as affective states and linguistic competencies, and external elements including rater behavior and contextual conditions plays a pivotal role in shaping the speaking exam performance. Both instructors and students agree on the crucial role of vocabulary and motivation, while also highlighting anxiety and readiness as significant hurdles. Focusing on these challenges as well as supporting the facilitating factors could lead to a more supportive and effective assessment environment, ultimately improving student outcomes in speaking proficiency. As such, implementing strategic changes to exam structures, enhancing clarity in task design, and fostering supportive examiner-student dynamics can be chief steps towards optimizing performance in speaking assessments for EFL learners.

5.2 Discussion

The findings of this study offer valuable insights into the complex nature of speaking performance in EFL achievement tests by triangulating instructors' and students' views via quantitative and qualitative data. The purpose of this section is to critically examine

the implications of these results with reference to existing literature and highlight possible implementations and improvements for oral assessments.

When the quantitative data were analyzed, the divergence between instructor and student perceptions regarding the primary factors affecting speaking performance became obvious. Instructors predominantly underlined linguistic competencies, especially vocabulary range and appropriacy beside motivation to speak; however, students focused more on task-related factors including number and variety of tasks as well as task type. This variation aligns with previous research conducted by Yule (2010) who highlights the importance of linguistic competence for effective communicative competence. Learners, on the other hand, bear speaking exams through a more situational perspective, which is in line with Lee's (2006) analysis of score dependability of a standardized English test, in which he pointed out the importance of increasing the number of tasks to enhance score dependability. Such findings indicate a need for consistency between teaching practices and assessment methods to better accommodate students' experiences and challenges. Thus, recognizing this perception gap could promote a more balanced perspective in the speaking curriculum that embraces both linguistic development and strategic task design. In the qualitative data, various factors emerged. Particularly, participants viewed affective factors such as anxiety and nervousness, students' readiness, linguistic competencies including accurate vocabulary and grammar use, content production, the effect of exam raters, specifically their behavior and familiarity with the students, and contextual elements as the most influential factors on students' speaking performance. The following section delves into these factors, respectively.

5.2.1 The role of affective factors

The role of affective factors, specifically anxiety, nervousness and self-perception emerged as the most notable themes. Participants described various forms of anxiety that hampered the ability to perform effectively. Learners often expressed concerns regarding their anxiety levels, supporting Horwitz et. al.'s (1986) foreign language anxiety theory, which posits that high anxiety levels can result in poorer speaking performance. In a similar vein, Goh and Burns (2012) pinpoint the adverse effects of high levels of anxiety or stress on learners' performance. The findings of Merç and Özsoy's (2022) study also highlight the negative effects of speaking anxiety on learners' willingness to communicate (WTC), which can affect learners' speaking performance. Likewise,

instructors in this study prioritize the learners' anxiety levels as a major determinant of their oral success. This supports the studies of Dewaele and MacIntyre (2019) and MacIntyre and Gregersen (2020) on the relationship between positive psychology and language anxiety, emphasizing that positive psychology reduces language anxiety and enhances speaking performance. The prominence of affective factors such as anxiety, and nervousness suggests that speaking performance is highly sensible to fluctuations in emotional state, which underlines the importance of affective readiness.

The existing literature also supports conforming findings with regard to the effects of affective factors on oral performance. To start with, in Hanifa's (2013) review, he tried to address the sources of speaking anxiety, and cognitive, affective and performance-related factors stand out. These affective factors include students' attitudes and emotions towards the topic and/or the interlocutors as well as self-consciousness. Likewise, as Woodrow (2012) suggests, speaking anxiety was found to be a significant predictor of oral achievement. Yilmaz (2015) examined the effect of test anxiety among Turkish EFL students, reflecting a strong correlation between test anxiety and low oral proficiency. Next, in Jahbel's (2017) study with 27 EFL high school learners and teachers, students indicated confidence and motivation, respectively, as the major hurdle in speaking. The study also revealed that half of the teachers stated the negative influence of anxiety on learners' speaking skills. Moreover, the results of Alrasheedi's (2020) study displayed the significant effect of affective factors including anxiety, shyness, fear of making mistakes and peer pressure on EFL university students' speaking test performance. In a similar vein, Amoah and Yeboah (2021) investigated the factors influencing the speaking performance of Chinese EFL learners, and the findings displayed psychological factors, including anxiety, fear of mistakes, unwillingness, and fear of negative evaluation, had a more significant impact on speaking performance than other examinee-related factors. Furthermore, Jaya et. al. (2022) investigated the relationship between EFL learners' speaking performance, speaking problems and the factors related to these problems. One of the major speaking problems addressed by the researchers is affect-related problems including confidence and anxiety. Finally, Dong (2022) investigated the factors affecting EFL students' speaking difficulties, specifically regarding linguistic, psychological, and environmental challenges. The findings revealed that psychological factors, particularly low self-confidence, posed the greatest difficulty.

On the other hand, there are studies that contradict the findings of this study. To exemplify, the analysis of Jahbel's (2017) study demonstrates that among 21 student participants, only a small number (33%) viewed anxiety and pressure to perform well as influential factors on their performance, while half of the teacher participants believed the role of anxiety on performance. This discrepancy might be attributed to the educational level of the participants, as Jahbel's (2017) study focused on high school learners, whereas the present study involved tertiary-level students. In addition, the cultural background of learners may have an influence on the amount of anxiety in L2 speaking. In the context of Türkiye, it can be seen that the society is composed of individuals from diverse cultural and national backgrounds, some of whom might display intense cultural sensitivity toward affective situations. Besides, Horwitz et. al. (1986), on the other hand, suggested a moderate amount of anxiety can have a facilitating effect, the type of anxiety referenced in Jahbel's study may function more as a motivational force rather than a hindering factor, which contrasts with the anxiety described by the students in the present study, who perceived it primarily as an obstacle to effective speaking performance.

Even though different results have been obtained, these studies underscore the noteworthy role of affective and psychological factors with some key emerging themes such as anxiety, fear of mistakes, and low self-confidence in EFL learners' speaking performance. The findings across different contexts are not only compatible but also unlimited to certain educational and cultural background, and thus, display the universal struggles among EFL learners. In the context in which this study was conducted, it is not surprising that the biggest factor affecting students' oral performance was emotional factors. Since it is a must for the participants of this study to be successful in preparatory class to be able to study in their majors, the influence of such affective variables could have been amplified by the academic pressures and performance expectations in tertiary education settings. Furthermore, as discussed by Merç and Özalp (2022), many students do not have the opportunity to gain proficiency in L2 communication prior to university education, which might be another reason for the high level of anxiety among students participated in this study. Therefore, the strong correlation between (test) anxiety and low oral proficiency highlights the need for educators, test designers and administrators to create supportive, low-stakes speaking environments. Addressing these psychological barriers

by incorporating anxiety-reducing strategies such as simulated test environments and peer-based preparation and confidence-building activities along with positive reinforcement can mitigate its negative impact on learners' performance and help them overcome their anxiety and eventually improve their speaking skills.

5.2.2 The role of readiness

In addition to affective factors, readiness, particularly practice in and outside the classroom was deemed as another crucial element in EFL speaking tests. Both instructors and students emphasized the beneficial effects of allocating time for practice, whereas the absence of sufficient practice was stressed as a key hurdle for oral success. This also points to speaking proficiency as a skill which must be continuously rehearsed rather than simply acquired through passive learning. Learners who engaged in more authentic and frequent speaking opportunities demonstrated greater fluency and confidence. DeKeyser (2007), for example, highlights the necessity of proceduralization through repeated practice for skill development. Likewise, according to Skehan (1998), students engaging in mock exams or who rehearsed speaking tasks had a tendency to exhibit lower anxiety and higher fluency. Similar to these findings, some research studies are also compatible with the existing literature. To exemplify, Jahbel (2017) investigated the factors that affect EFL high school students' speaking performance with 21 students and six teachers via questionnaires and interviews. The analysis of the results displayed various factors that affect their performance, yet all of the teacher participants agreed on the "time for preparation" and "motivation" as the most impactful elements on students' speaking performance. Similarly, 57% of the students indicated the negative effects of time for preparation. Second, Qasemi (2020) examined the factors that affect students' speaking performance in an English department, and identified certain difficulties caused by a lack of opportunity to practice outside the classroom. Similarly, Jaya et. al. (2022) noticed that more than half of the participants did not perform well in the speaking test, so they conducted research on the effects of EFL university students' speaking performance, speaking problems, and factors associated with these problems. The results yielded various reasons of poor performance, among which lack of speaking practice was conspicuous. Furthermore, though Alrasheedi (2020) emphasized the significant role of affective factors, the results of the study also yielded the lack of exposure to English and insufficient out-of-class speaking practice opportunities as other struggles for learners.

Also, the participants in Dong's (2022) study stated lack of opportunities to practice out-of-class speaking, emphasizing environmental constraints. Finally, in his study, Liu (2022) highlighted the importance of learners' discussion abilities; thus, he developed a program called Schema Association Training (SAT) to assist his learners to improve their discussion abilities in and outside the classroom as well as boost their confidence during speaking assessment.

Additionally, in the current study, a well-understood lesson and the ability to recognize question patterns or knowing the exam procedure also helped students feel more confident. These factors echo the findings of Shohamy (2001), who emphasizes the enhancing effects of familiarity with the assessment format on students' test-taking confidence and decreased anxiety.

These aforementioned studies highlight the significant effect of readiness on EFL students' speaking performance, as it can either enhance their success or make oral production a real challenge. The importance attached to practice within and beyond the classroom may depend on learners' cultural values. In some cultures, it is known that academic diligence is hugely emphasized, which may bring on greater importance to individual study and consistent practice. The fact that participants of this study emphasized the importance of practicing not only in the classroom but also outside the classroom, and even made self-criticism in this regard, in so doing, once again underlines the importance of readiness. It is clearly understood that learners who had opportunities to rehearse or engage with similar tasks beforehand demonstrated improved performance. Therefore, creating opportunities for learners to engage in extensive speaking practice both within and beyond the classroom becomes pivotal for fostering their oral abilities. Language teachers could incorporate in-class interactive speaking activities like role-plays and discussions to encourage students to practice in a structured yet communicative environment. Furthermore, integrating technology via language learning apps and online conversation partners could afford further opportunities for out-of-class practice. By applying these particular strategies, teachers may assist students build confidence, reduce anxiety, and eventually improve their overall oral skills.

5.2.3 The role of linguistic competencies

Another major factor cited in the qualitative data was related to learners' linguistic competencies. A considerable number of participants explicitly mentioned the importance

of language skills, notably vocabulary skills. This is also in line with the previous research since many scholars emphasize the need of linguistic competence for improved oral communication skills (Bailey, 2020; Ellis, 2008; Harris, 1990; Nunan, 1991; Ur, 1996; Yule, 2010). For example, Al Hosni (2014) found that the majority of the participants in his study regarded speaking in English as a challenging task because of having limited vocabulary and grammar structures as well as mother tongue use. Similarly, Afeabri and Muhsin (2019) concluded that among various factors, linguistic factors such as grammar, vocabulary and pronunciation were ranked the highest regarding the factors that affect 10th-grade EFL learners' speaking exam performance. Third, in Jaya et. al.'s (2022) study, besides a lack of speaking practice, linguistic problems related to vocabulary, grammar, pronunciation and fluency were found to be important reasons that led to speaking problems. Additionally, although Alrasheed's (2020) study focused on affective factors, the results also revealed that limited vocabulary was determined as another inhibitory factor. Similar to Alrasheed's (2020) study, Dong (2022) explored the factors affecting EFL students' speaking difficulties, specifically regarding linguistic, psychological, and environmental challenges. Linguistic issues such as limited vocabulary for expressing ideas and arguments were found to be the second major element after psychological factors. In another study conducted by Saka and Merç (2021), EFL learners perceived themselves as having enough self-efficacy. However, those who reported lower speaking efficacy explained that their lack of vocabulary made them feel less efficient.

Contrary to all these finding, Amoah and Yeboah's (2021) research indicated that linguistic factors such as limited vocabulary, pronunciation difficulties, and insufficient grammar knowledge had less impact on Chinese tertiary level EFL students' speaking performance when compared to psychological factors, including anxiety, fear of mistakes, unwillingness, and fear of negative evaluation. Though the participants in Amoah and Yeboah's (2021) study involve tertiary level EFL learners, the reason for this difference may be due to cultural variances. In cultures such as traditional Chinese culture, emotional factors can take precedence over linguistic factors when speaking a second language.

The findings of this study align with previous research encompassing the interplay between linguistic competence and speaking proficiency. Limited vocabulary, grammatical inaccuracies, and pronunciation difficulties can hinder fluency and

confidence, making speaking difficult for EFL learners. As Fulcher (2014) states, test-takers are expected to produce accurate, appropriate and intelligible speech utterances, which results in pressure on learners as speaking under time pressure. Considering that vocabulary and grammatical competence are among the keystones of speaking ability, this finding is very relevant. The importance attributed to this factor might stem from the expectations of stakeholders regarding linguistic competencies during L2 oral production. It also indicates that linguistic readiness is as critical as emotional and practice-based readiness in oral assessments. Hence, educators can consider applying focused instructional strategies like explicit vocabulary instruction, grammar-focused speaking tasks and pronunciation practice to address these difficulties. Also, helping learners overcome the issue of lack of essential skills through communicative approaches, which allow them to utilize their linguistic knowledge in meaningful contexts, could create opportunities for more natural and effective language use.

5.2.4 The role of content production

A further key aspect emerging in the qualitative data was about content production. Cognitive processing during speaking requires more than accessing previously stored knowledge; that is, it requires flexible and rapid idea organization under pressure. Thus, generating relevant and immediate content can be a challenge for students. Especially instructors regarded learners' inability to generate ideas as a major challenge, resulting in limited content production. This finding aligns with Tuan and Mai's (2015) investigation on the speaking problems of Vietnamese EFL learners, in which the results of a teacher questionnaire indicated that students' major problem in speaking was the lack of topical knowledge. Moreover, Jahbel (2017) found that 66% of teachers and 57% of students viewed that topical knowledge influenced students' oral performance negatively. Huang et. al. (2018) also investigated the impact of topical knowledge on independent and integrated speaking test performance, in which the results yielded the significance effect of topical knowledge on integrated speaking test performance. Likewise, Jaya et. al.'s (2022) study displayed a lack of general knowledge as a factor of speaking problems among EFL learners.

These findings suggest that beyond linguistic proficiency, helping learners develop ability to generate ideas is crucial for enhancing their oral communication skills. In a face-to-face speaking test, it is obvious how difficult it is to come up with ideas instantly

alongside nervousness. Even learners with high linguistic competence struggled when they lacked ideas or failed to develop responses meaningfully as content development requires cognitive skills. This might stem from not only emotional and linguistic elements, but also from lack of certain habits such as reading. In Türkiye, unfortunately, the rate of engagement with various forms of reading materials remains significantly low, including among university students. This can be one of the major obstacles to content production. Thus, incorporating activities such as extensive reading tasks or reading circles into curriculum and adding extra-curricular tasks that can reinforce critical thinking, brainstorming, and knowledge-building may help learners tackle the issue and contribute to more fluent and meaningful verbal expression.

5.2.5 The role of exam raters

An additional noteworthy factor arising from the qualitative findings is based on the role of exam raters. Both participant groups highlighted the effects of rater behavior and rater familiarity as significant aspects of oral success. To be precise, both instructors and learners consider the friendly demeanor of raters and their acquaintance with them as significant factors, which may unintentionally advantage or disadvantage the learners during the evaluation process. This finding supports McNamara's (1996) argument that rater characteristics, such as harshness or leniency, can introduce variability in speaking assessment outcomes. To be more specific, Winke et al. (2012), for example, found that assessors who spoke the same native language as the test takers were more lenient in scoring. Eckes (2015), for another example, explored rater effects and biases in assessing German EFL learners' performance by analyzing rater scoring patterns and interviews, and highlighted the influence of individual raters' characteristics on speaking test outcomes. In addition, Neittaanmäki and Lamprianou (2024) examined the changes in rater severity and consistency with Finnish raters. Findings of the study revealed that raters became slightly more lenient over time. Considering that experienced teachers are generally perceived as more understanding and lenient than the less experienced teachers, this finding is consistent with both existing literature (Neittaanmäki & Lamprianou, 2024) and my own professional observations and experiences within educational settings.

In addition, although mentioned by a relatively small number of participants, the way examiners speak and teachers' maximum use of English in class were also regarded as facilitators of better performance. Consistent use of model target language and providing

students with richer linguistic input, supporting Krashen's (1982) Input Hypothesis, posits that language acquisition flourishes with abundant comprehensible input. Furthermore, this factor may be influenced by the teachers' level of experience. Studies not previously discussed but relevant to interpreting this outcome are those of Borg (2006) and Tsang and Chow (2011), which examine the practices of novice and experienced teachers. The findings of these studies yield that novice teachers are often observed to demonstrate higher levels of enthusiasm and motivation when using the target language in the classroom compared to their more experienced counterparts, who may adopt a more routine-based or procedural approach to language instruction. Considering the number of experienced instructors teaching at the institution where the data were gathered, it is reasonable to assume that teaching experience may have played a role in shaping the observed outcomes.

The findings put emphasis on the importance of rater influence as a critical factor in assessing oral performance. In the context in which the study was carried out, although a significant number of students mentioned the positive effects of knowing the testers in advance, rater familiarity may result in potential bias, too. For that reason, it might be ideal for the examiners to adopt a positive and non-biased attitude during the speaking exams so that students may feel a sense of comfort in a process that is already stressful enough. A friendly and supportive demeanor could alleviate anxiety and provide a more enabling context for students to perform at their best. Furthermore, examiners familiarizing themselves with the test-takers prior to the exam performance may contribute to a feeling of security and ease. This can lead to fair assessment of learners' speaking proficiency, as well.

5.2.6 The role of contextual factors

Contextual elements of speaking tests including peer effect, task questions, exam conditions such as having time to think on the exam question and physical context like video-recording were derived as final factors. These environmental and procedural factors can be often considered peripheral, yet they may play a central role in determining learners' oral abilities. Of contextual factors, peer effect including unfamiliar exam partners and negative comments of students who took the exam previously stands out. Vygotsky's (1978) sociocultural theory supports these findings, proposing that collaborative interaction within the Zone of Proximal Development (ZPD) promotes

higher levels of performance. That is, learners acquire the target language when they are in a more collaborative environment, in which peers can support one another. Similarly, Young (1991) suggests speaking in the presence of peers has been shown to trigger evaluative apprehension. Likewise, in their study which examined the relationship between Turkish EFL learners' WTC in English, self-efficacy and linguistic self-confidence, Saka and Merç (2021) concluded that EFL students had higher WTC in English, especially when presenting a talk to a group of friends rather than a group of strangers. However, to my humble mind, peer effect on the speaking tests has not been probed in the literature in detail. One study conducted by Kwon (2015) investigated the effects of paired speaking on Korean middle school EFL learners in a mixed methods research design, in which the findings mainly reflected a positive perception towards the paired format of the speaking exam since they could speak with confidence. Nonetheless, some students deemed pair format tests as pressure and anxiety evoking as they felt it decreased the motivation and increased the competitiveness. These results are in line with the findings of this current study, indicating that students reported feeling more at ease when they were with a familiar exam partner. This might be attributed to the opportunities to practice collaboratively if they were in the same classroom, which allowed them to observe each other's strengths and weaknesses so that they could mutually support one another during the exam. Whilst peer presence occasionally amplifies anxiety, it also gave students support to perform better. On the contrary, exposure to negative comments from previous test-takers was associated with high level of anxiety by some participants.

Task questions arose as another context-based factor. Specifically, participants emphasized that they were challenged by the tasks when the topics in the questions were new to them or the format of the exam questions (e.g., discussion) was difficult. On the other hand, some students favored picture-prompted questions as they aided in generating ideas. According to Fulcher (2003), complex or unfamiliar task formats can impair students' fluency and confidence. This also aligns with O'Grady's (2019) research, whose results indicated increases in test-takers' scores on the picture-based narrative tasks. Likewise, the finding of another study by Hsiegh and Wang (2017) showed that task types such as picture narration or integrated listen/speak tasks had some effects on learners' linguistic elements such as grammar, vocabulary and content. Additionally, the findings gathered in Lee's (2006) study underscored the importance of the number of tasks as it

was found to be more effective in developing score dependability among raters. Üstün's (2010) research also suggested integrating authentic tasks so that students' performance and engagement could be improved. Another suggestion was made by Crossley and Kim (2019), whose study results indicated text integration as a significant factor on participants' speaking proficiency scores. In this study, tasks perceived as vague or unfamiliar were reported to contribute to increased hesitation and low speaking performance. How the question is asked, rather than what it asks, can be a condition for completing the task. For example, in this achievement test where the data were collected, students may not have displayed their actual performance because of some question types that used pictures or required interaction. They might not be able to make connections between the question and the pictures, or may not be able to initiate and maintain a conversation with a partner in the exam.

Exam conditions involved elements such as non-stressful environment, distribution of the questions and thinking time. Although there is not much empirical research on these issues in the literature, and were cited by a small number of participants in this present study, the existence of these factors cannot be ignored. Bachman (2004) underlines the importance of creating a supportive and conducive testing environment where students are given enough time to think and respond. This can, by extension, enhance the validity and reliability of the speaking assessments. Consistent with O'Grady's (2019) study, the duration of thinking time given to test-takers in the exam, the analysis showed that students who were given 5 minutes for planning and organizing their speech, displayed better overall results in their oral tests. This finding aligns with the finding in this study as a student felt having thinking time as a facilitator in their oral performance. This might be due to some instructors' tendency to be impatient and initiate the exam without allowing students to fully utilize the allotted one or two-minute thinking time, despite its being part of the exam procedure.

Final context-based factors are associated with the physical context of the exam room such as video-recording, exam room environment like seating, lighting, etc., multiple examiners and slot. Similar to exam conditions, the literature lacks relevant research. Even so, according to O'Sullivan (2000), external distractions can be reduced with a well-structured, physically comfortable exam environment, allowing students to focus more on their cognitive and linguistic tasks. However, in this context, the narrow corridors of

the institution might have contributed to noise generated by test-takers who had previously completed their exams, potentially affecting the performance of those still being assessed. With regard to video-recording, the findings of this study demonstrated that one-third of the teachers regarded it as a challenge; however, it was cited by only one student. This difference might be the result of the generation Z's having been grown up in an environment with constant camera presence. A study that is relatively close to this result, conducted by Alzahrani (2020), can be mentioned. In his study, he compared two modes of oral proficiency testing: Oral Proficiency Interview (OPI) and Video Oral Communication Instrument (VOCI) regarding test-takers' complexity, accuracy and fluency. The accuracy and the fluency level were found higher in OPI, and was favored by the participants since they felt more comfortable with this testing mode. The reason might be due to uncomfortable feelings derived from being in front of a video.

Although the aspects concerning exam conditions and contextual factors were mentioned by a limited number of participants and remain underexplored in the existing literature, their potential influence on test-takers' performance should not be overlooked. Some participants in this study stressed elements such as a non-stressful environment, distribution of questions to learners, and the inclusion of thinking time as influential in shaping their speaking performance outcome. Furthermore, contextual elements related to the physical exam environment such as seating, lighting, presence of video recording, number of examiners, and exam scheduling were also identified as relevant, though underexplored, contributors to the overall test experience. Therefore, the test designers and the administrators need to be aware of these subtle aspects of exam settings, and the discomfort stemming from exam context such as being recorded. By and large, it would be wise for the testers to arrange the exam setting accordingly, and create a stress-free environment in which negatively influencing factors are eliminated, especially in high-stakes speaking tests. All in all, testing environments can have subtle but vital impacts on performance, specifically in spoken tasks that require concentration and emotional regulation; thus, it is crucial to understand the role of these factors on learners' speaking performance and find ways to overcome these challenges in EFL settings.

5.3 Pedagogical Implications

The purpose of this study is to find out the factors that affect EFL learners' speaking performance in achievement tests. Findings of the study shed light on a range of

pedagogical implications regarding some themes that emerged to be of utmost importance for EFL language instructors, curriculum developers and test designers seeking to enhance EFL learners' speaking performance in achievement tests. Based on these themes, the following implications can be drawn.

A central outcome of the study is the identification of several interrelated elements affecting learners' speaking outcomes. These elements that impact learners' performance include affective factors, readiness for the exams, and linguistic factors including language proficiency and content production ability. In particular, the prominence of affective factors such as anxiety, fear of making mistakes, and lack of readiness appear to hinder students' ability to express themselves verbally. These internal factors show that speaking performance is closely bound to learners' readiness and psychological state during assessments, making the affective domain a critical aspect of language performance.

Another critical implementation concerns getting ready for these oral exams. Consistent and meaningful opportunities for spoken language practice are essential for both improving oral fluency and for helping learners become more confident and less susceptible to negative emotions during assessments. Therefore, extensive speaking practice in regular classroom instruction becomes significant not only for both language development but also for minimizing the adverse effects of affective and psychological factors. In such, the value of social interaction tasks including turn-taking at talk should not be underestimated since tasks including social interaction can better lead to authenticity and real-world use of language.

Moreover, the linguistic challenges particularly limited vocabulary and difficulties in content generation, further impact learners' performance. The inability to retrieve or use appropriate language under exam conditions highlights not only the lack of linguistic knowledge but also the cognitive load learners experience during speaking tasks. This points out that language retrieval during speaking tests is not only a matter of knowledge but also of processing fluency under pressure.

Another noteworthy implication relates to the role of the examiner. Learners appeared to perform more effectively when interacting with the examiners who are supportive, empathetic, positive and familiar to the examinees. This highlights the importance of interpersonal dynamics and rapport in a speaking exam environment, and points to how

these dynamics either ease or aggravate negative feelings between examinees and examiners, which may vary across cultural contexts in which perceptions of authority, familiarity, and emotional expression can differ significantly.

Lastly, the influence of peer presence, especially the presence of familiar peers during the exam, may appear to influence learners' anxiety and comfort levels. Some students felt more at ease and performed better when paired or grouped with familiar peers. Educators may consider grouping students thoughtfully or allowing familiar pairings in certain oral assessments to reduce anxiety and increase comfort.

When everything is taken together, these findings underline that EFL speaking performance in testing contexts is a multi-dimensional construct which is shaped by affective, linguistic, interpersonal, and contextual variables. Addressing these dimensions holistically in pedagogical and assessment practices may lead to more valid and equitable evaluations of learners' speaking proficiency. To this end, the following section will outline a number of recommendations aimed at boosting learners' oral performance during the examinations.

5.4 Suggestions

In light of the findings, several practical suggestions can be made for the EFL educators, curriculum designers, and test developers as well as directions for future research.

5.4.1 Suggestions for EFL educators

The first, and perhaps most important, suggestion encompasses the role of EFL teachers in shaping the learning atmosphere. EFL educators should create a friendly and non-threatening (Merç & Özsoy, 2022), as well as psychologically safe and culturally appropriate learning and testing environment. To achieve this, teachers could implement anxiety-reduction techniques such as relaxation strategies, and positive feedback approaches because incorporating low-stakes speaking activities and providing regular constructive feedback may help learners build self-confidence and reduce performance anxiety.

Secondly, teachers can design in-class tasks that require social interaction, as well. Some types of tasks depending on real-world use may better lead to the authentic use of language, particularly, when tasks include social interactional requirements, such as turn-taking, negotiation of meaning, and collaborative problem-solving. These interaction-

focused tasks can both motivate learners to practice spontaneous speaking and also help them develop pragmatic competence and conversational skills necessary for effective communication.

In addition to in-class tasks, the learners need more practice beyond the classroom. Therefore, the teachers could design enhanced in-class tasks that can also be extended outside the classroom so that students can have more extracurricular speaking activities that can lead them practice outside the classroom. These out-of-class activities can take the form of a digital portfolio of speaking tasks, or the form of tasks that require different mobile instruments like apps or platforms such as Duolingo, ELSA Speak, Lingbe. In addition, some chatbots powered by artificial intelligence (AI) such as ChatGPT (Voice mode) and TalkPal AI can be utilized to develop speaking skills. These tools can offer learners flexibility in terms of place, time, pace, and the choice to work individually or with peers, thereby fostering a low-anxiety environment for practicing speaking skills.

Next, since learners struggle with language specifically vocabulary and content production, integrating lexical development into regular speaking practice is essential. Structured speaking practice sessions emphasizing vocabulary building and idea generation can prepare learners for the oral tests more effectively. Teachers can also encourage the use of speaking outlines or brainstorming techniques before performance tasks. Moreover, teachers can prepare more speaking tasks that focus on fluency and idea development rather than accuracy alone. Additionally, instructors can explicitly teach certain strategies such as lexical chunking, thematic vocabulary review, and idea scaffolding that may help ease the cognitive demands of speaking assessments.

5.4.2 Suggestions for curriculum designers

Besides suggestions for EFL instructors, several important suggestions can be offered to the EFL curriculum designers. First of all, to reduce test anxiety and fear of making mistakes, more pre-test practice activities and confidence-building exercises such as role-plays and interviews should be incorporated into the existing curriculum. Though recorded mock exams were shown in the classrooms before the speaking tests, clearly, there was still need to implement these exams in the classrooms as if they were in a real exam environment. Designing familiar and authentic tasks and exams with familiar, real-world topics as well as topics covered in the classes can help lower anxiety and boost performance.

As mentioned above, the learners need more practice not only in but also outside the classroom. Thus, along with classroom instructors, curriculum designers should incorporate such speaking tasks into the existing curriculum. Where possible, curriculum designers can facilitate opportunities for students to interact with native speakers or proficient non-native speakers. This can encourage students to use the maximum amount of the target language in authentic communication contexts. In brief, helping students be engaged in real-life related language tasks, and make it a part of their own lives become essential.

5.4.3 Suggestions for test developers

To address the role of examiners, particularly regarding their demeanor and familiarity to students, test developers at institutions should organize training workshops that emphasize supportive interaction and awareness of their potential impact on student performance. Also, training and norming sessions should be supported to maintain consistency and to minimize rater-related variability in assessment. These trainings can also encompass how to maintain a supportive tone and clear guidance. Additionally, it may be beneficial to keep examiner-learner consistency across speaking test administrations in an academic year if possible. In this way, test-takers can be acquainted with the accent and the way the testers speak, minimizing rater-related elements during assessment.

In terms of task design, test designers might use familiar and personally relevant topics as best they can. Secondly, exam questions can be displayed in written form beside orally announcing to the students to avoid ambiguity or difficulty in understanding the exam questions due to remote reading. In addition, multimodal prompts such as pictures can help activate students' prior knowledge and, hence, lessen cognitive load during the exam. Furthermore, for lower-level learners, questions requiring pair interaction may be omitted, as authentic interaction entails more complex skills that may exceed the learners' current proficiency level. Finally, although the presence of peers can cause anxiety for some, allowing learners to choose or become familiar with their speaking partners before the exam could help foster a more relaxed atmosphere. Thus, while arranging the exam pairs, the students from the same classroom should be paired up to reduce the stress stemming from an unknown exam partner.

5.4.4 Suggestions for future research

In addition to the suggestions to language educators, curriculum designers, and test developers, it would also be appropriate to include some suggestions for future studies, as well. To start with, further research could explore how these variables interact across different age groups and cultural contexts. That is, since the participants were all Turkish EFL learners studying at a state university's preparatory classes, various learners with different age groups and backgrounds can participate in future research. Considering that speaking exams have also been incorporated into the national curriculum by the Ministry of National Education, it would be meaningful to broaden the scope of this study, and investigate these factors in public middle and high schools where speaking assessments are implemented.

Likewise, to better understand the phenomenon, the sample size could be expanded by gathering data from different universities in which speaking exams are implemented. Also, longitudinal studies could examine whether interventions such as anxiety-reduction strategies or task design changes lead to measurable improvements in oral performance over time. In addition, future studies might investigate examiner-related factors more deeply, such as the impact of examiner training or familiarity in high-stakes testing environments.

Finally, the present study investigated the factors that affect EFL learners' oral performance only in achievement tests. Its scope could be further expanded to include factors in other types of tests such as proficiency tests. Additionally, future research can investigate the comparisons between human scoring and AI-based scoring systems in speaking assessments. Such a comparison may reveal potential facets in, for example, scoring consistency, rater bias, and feedback quality, which can lead to the contribution to the development of fairer and more reliable assessment methods.

All in all, addressing the multi-dimensional and complex nature of the factors influencing EFL learners' speaking performance requires a holistic approach, comprising both internal and external elements of the testing experience. By acknowledging the affective, linguistic, and contextual dimensions revealed in this study, language educators, assessment developers, curriculum designers and institution administrators may collaborate toward creating more equitable and supportive oral assessment environments. Thoughtful attention to these aspects not only improves test performance but also

contributes to learners' long-term communicative competence and self-confidence in real-world speaking situations.



REFERENCES

- Afebri, H. A., & Muhsin, K. (2019). Factors affecting students' difficulties in speaking performance of the tenth grade students of SMA Negeri 1 Tiworo Kepulauan. *Journal of English Education and Teaching*, 3(1), 95–105. <https://doi.org/10.33369/jeet.3.1.95-105>
- Ahmadi, A., & Sadeghi, E. (2016). Assessing English language learners' oral performance: A comparison of monologue, interview, and group oral test. *Language Assessment Quarterly*, 13(4), 341–358.
- Akbari, R. (2012). Validity in language testing. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoyanoff (Eds.), *The Cambridge guide to second language assessment* (pp. 30–36). Cambridge University Press.
- Al Hosni, S. (2014). Speaking difficulties encountered by young EFL learners. *International Journal on Studies in English Language and Literature (IJSELL)*, 2(1), 22–30.
- Alonso, R. A. (Ed.). (2018). *Speaking in a second language* (Vol. 17). John Benjamins Publishing Company.
- Alrasheedi, S. (2020). Investigation of factors influencing speaking performance of Saudi EFL learners. *Arab World English Journal*, 11(1), 1–20. <https://dx.doi.org/10.24093/awej/vol11no1.1>
- Alzahrani, N. A. (2020). *A comparative study of oral proficiency in direct (OPI) and semi-direct (VOCI) testing modes: Measures of complexity, accuracy, and fluency* (Doctoral dissertation, Oklahoma State University). ProQuest Dissertations Publishing.
- Amoah, S., & Yeboah, J. (2021). The speaking difficulties of Chinese EFL learners and their motivation towards speaking the English language. *Journal of Language and Linguistic Studies*, 17(1), 56–69.
- Babaii, E., Taghaddomi, S., & Pashmforoosh, R. (2016). Speaking self-assessment: Mismatches between learners' and teachers' criteria. *Language Testing*, 33(3), 411–437.
- Bachman, L. F. (1990). Fundamental considerations in language testing. *Oxford University Press*.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.
- Bailey, K. M. (2020). Teaching listening and speaking in second and foreign language contexts. *Bloomsbury Academic*.
- Boers, F., Eyckmans, J., & Stengers, H. (2020). Evidence for the benefits of a phrase-noticing pedagogy. *Language Teaching Research*, 24(3), 345–369.
- Borg, S. (2006). The distinctive characteristics of foreign language teachers. *Language Teaching Research*, 10(1), 3–31.
- Brindley, G. (1991). Defining language ability: The criteria for criteria. In S. Anivan (Ed.), *Current developments in language testing* (pp. 139–164). *Regional Language Centre*.
- Brown, H. D. (2003). Language assessment: Principles and classroom practices. *Pearson Education*.
- Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices* (2nd ed.). Pearson Education.

- Brown, A., Iwashita, N., & McNamara, T. (2005). An examination of rater orientations and test taker performance on English for academic purposes speaking tasks. *ETS Research Report Series*, 2005(1), i–157. <https://doi.org/10.1002/j.2333-8504.2005.tb01995.x>
- Brown, J. D. (2015). Mixed methods research. In C. Coombe, S. J. Stoyonoff, P. Davidson, & B. O'Sullivan (Eds.), *The Cambridge guide to research in language teaching and learning* (pp. 78–84). *Cambridge University Press*.
- Burch, A. R., & Kasper, G. (2021). Task instruction in OPI roleplays. In M. R. Salaberry & A. R. Burch (Eds.), *Assessing speaking in context: Expanding the construct and its applications* (pp. 63–86). *Multilingual Matters*.
- Burstein, J., Chodorow, M., & Leacock, C. (2013). The efficacy of automated scoring for speaking tests. *Journal of Educational Measurement*, 50(2), 132–154.
- Bygate, M. (1987). *Speaking*. Oxford University Press.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1–47.
- Chalhoub-Deville, M., & Wigglesworth, G. (2005). Rater judgment and English language speaking proficiency. *World Englishes*, 24(3), 383–391.
- Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Pearson Education.
- Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). SAGE Publications.
- Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications.
- Creswell, J. W., & Guetterman, T. C. (2019). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (6th ed.). Pearson.
- Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications
- Crossley, S. A., & Kim, Y. (2019). Text integration and speaking proficiency: Linguistic, individual differences, and strategy use considerations. *Language Assessment Quarterly*, 16(2), 217–235.
- DeKeyser, R. M. (2007). *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*. Cambridge University Press.
- Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, 39(3), 379–397.
- Dewaele, J.-M., & MacIntyre, P. D. (2019). The predictive power of multicultural personality traits, learner and teacher variables on foreign language enjoyment and anxiety. In M.
- Doff, A. (1998). *Teach English: A training course for teachers*. Cambridge University Press.
- Dong, M. H. (2022). A study on factors affecting of English-majored students' difficulties in their speaking performance. *European Journal of Foreign Language Teaching*, 6(1).
- Dörnyei, Z. (2007). *Research methods in applied linguistics*. Oxford University Press.
- Dörnyei, Z., & Ryan, S. (2015). *The Psychology of the Language Learner Revisited*. Routledge.

- Eckes, T. (2015). Examining rater effects in language testing. In G. Fulcher & F. Davidson (Eds.), *The Routledge Handbook of Language Testing* (pp. 364-376). Routledge.
- Ellis, R. (2003). *Task-based Language Learning and Teaching*. Oxford University Press.
- Ellis, R. (2008). *The Study of Second Language Acquisition* (2nd ed.). Oxford University Press.
- Ellis, R. (2017). *Position paper: Moving task-based language teaching forward*. *Language Teaching*, 50(4), 507-526.
- Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). SAGE Publications.
- Fulcher, G. (2000). The 'communicative' legacy in language testing. *System*, 28(4), 483-497.
- Fulcher, G. (2003). *Testing second language speaking*. Pearson Education.
- Fulcher, G., and Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London: Routledge.
- Fulcher, G. (2014). *Testing Second Language Speaking*. Routledge.
- Fulcher, G., & Davidson, F. (2016). *The Role of Rating Scales in Speaking Assessments*. Routledge.
- Gardner, R. C. (1985). *Social Psychology and Second Language Learning: The Role of Attitudes and Motivation*. Edward Arnold.
- Garrison, C., & Ehringhaus, M. (2007). *Formative and summative assessments in the classroom*. AMLE.
- Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Lawrence Erlbaum Associates.
- Goh, C. (2007). *Teaching speaking in the language classroom: Choices and issues*. SEAMEO Regional Language Centre.
- Goh, C., & Burns, A. (2012). *Teaching speaking: A holistic approach*. New York: Cambridge University Press.
- Gronlund, N. E. (1998). *Assessment of student achievement*. Allyn & Bacon.
- Gulikers, J. T., Bastiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational technology research and development*, 52(3), 67-86.
- Han, Q. (2016). Rater cognition in L2 speaking assessment: A review of the literature. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, 16(1), 1-24.
- Han, Q. (2020). *Investigating the combined effects of rater expertise, working memory capacity, and cognitive functionality on the scoring of second language speaking performance* (Publication No. 28022444) [Doctoral dissertation, Teachers College, Columbia University]. ProQuest Dissertations & Theses Global. <https://www.proquest.com/dissertations-theses/investigating-combined-effects-rater-expertise/docview/2427494755/se-2?accountid=7181>
- Hanifa, R. (2018). Factors generating anxiety when learning EFL speaking skills. *Studies in English Language and Education*, 5(2), 230-239.

- Harlen, W., & James, M. (1997). Assessment and learning: differences and relationships between formative and summative assessment. *Assessment in education: Principles, policy & practice*, 4(3), 365-379.
- Harmer, J. (2001). *The practice of English language teaching*. (3rd ed.). Longman.
- Harris, D. P. (1969). *Testing English as a Second Language*. McGraw-Hill Book Company.
- Heaton, J. B. (1975). *Writing English language tests: A practical guide for teachers of English as a second or foreign language* (10th ed.). Longman Publishing Group.
- Heaton, J. B. (1990). *Classroom testing*. Longman.
- Hedge, T. (2001). *Teaching and learning in the language classroom* (Vol. 106). Oxford, UK: Oxford University Press.
- Heigham, J., & Croker, R. (Eds.). (2009). *Qualitative research in applied linguistics: A practical introduction*. Palgrave Macmillan.
- Henning, G. (1987). *A guide to language testing: Development, evaluation, research*. Newbury House Publishers.
- Hsieh, C. N., & Wang, Y. (2019). Speaking proficiency of young language students: A discourse-analytic study. *Language Testing*, 36(1), 27-50.
- Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal*, 70(2), 125-132.
- Howatt, A. P. R., & Widdowson, H. G. (2004). *A History of English Language Teaching* (2nd ed.). Oxford University Press.
- Huang, H. T. D. (2016). Test-taking strategies in L2 assessment: The test of English for international communication speaking test. *Perceptual and Motor Skills*, 123(1), 64-90.
- Hughes, A. (2020). *Testing for language teachers*. Cambridge University Press.
- Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing*, 25(3), 385-402.
- In'nami, Y., & Koizumi, R. (2016). Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. *Language testing*, 33(3), 341-366.
- Isaacs, T., & Trofimovich, P. (2019). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Linguistics*, 40(2), 396-416.
- İnan, T. (2019). *Rater Cognition in the Assessment of EFL Speaking Performance* (Doctoral dissertation, Marmara University (Turkey)).
- Jaya, H. P., Petrus, I., & Pitaloka, N. L. (2022). Speaking performance and problems faced by English major students at a university in South Sumatera. *Indonesian EFL Journal*, 8(1), 105-112. <https://doi.org/10.25134/ieflij.v8i1.5603>
- Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed Methods Research: A Research Paradigm Whose Time Has Come. *Educational Researcher*, 33(7), 14-26.
- Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(1), 130-144.

- Joshi, A., Kale, S., Chandel, S., & Pal, D. K. (2015). Likert Scale: Explored and explained. *British Journal of Applied Science & Technology*, 7(4), 396–403.
- Kaçar, K. (2023). The purpose and principles of testing in ELT. *Alanyazın*, 4(1), 33–41.
- Karslı, E. S. (2002). *The inter-rater reliability of two alternative analytic grading scales for the evaluation of oral interviews at Anadolu University School of Foreign Languages* (Master's thesis, Bilkent University, Turkey).
- Khabbazzbashi, N., & Galaczi, E. D. (2020). A comparison of holistic, analytic, and part marking models in speaking assessment. *Language testing*, 37(3), 333-360.
- Kitao, S., & Kitao, K. (1996). Testing speaking (ERIC Document No. ED398261). ERIC. <https://eric.ed.gov/?id=ED398261>
- Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating scale developers. *Language Testing*, 28(1), 1-19.
- Knoch, U., & Elder, C. (2013). A framework for testing speaking. In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (Vol. 4, pp. 1-16). Wiley.
- Krashen, S. D. (1982). *Principles and Practice in Second Language Acquisition*. New York: Pergamon Press.
- Kumaravadivelu, B. (2006). *Understanding Language Teaching: From Method to Postmethod*. Lawrence Erlbaum Associates.
- Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. Longman.
- Lantolf, J. P. (2005). Sociocultural and second language learning research: An exegesis. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 335–353). Mahwah, NJ: Lawrence Erlbaum Associates
- Larsen-Freeman, D., & Anderson, M. (2011). *Techniques and Principles in Language Teaching* (3rd ed.). Oxford University Press.
- Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. MIT Press.
- Levis, J. M. (2018). *Intelligibility, Oral Communication, and the Teaching of Pronunciation*. Cambridge University Press.
- Lightbown, P. M. & Spada, N. (2021). *How Languages Are Learned*. 5th Edition. Oxford University Press.
- Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. SAGE Publications.
- Liu, G. (2022). Schema Association Training: Empowering Students for Speaking Assessments. *RELC Journal*, 53(1), 216-224.
- Long, M. H. (1980). *Input, interaction, and second language acquisition*. University of California, Los Angeles.
- Luoma, S. (2009). *Assessing speaking*. Cambridge: Cambridge University Press.
- Lyle, J. (2003). Stimulated recall: A report on its use in naturalistic research. *British Educational Research Journal*, 29(6), 861–878.
- MacIntyre, P. D., & Gregersen, T. (2020). Positive psychology in SLA. *System*, 91, 102262.
- Madsen, H. S. (1983). *Techniques in Testing*. Oxford University Press, 200 Madison Ave., New York, NY 10016 (ISBN-0-19-434132-1, \$5.95).

- McNamara, T. (1996). *Measuring second language performance*. Longman.
- McNamara, T., & Roever, C. (2006). *Language Testing: The Social Dimension*. Blackwell Publishing.
- Merç, A., & Özalp, Ö. S. (2022). The relationship between speaking anxiety and willingness to communicate: A quantitative inquiry. *Anadolu University Journal of Education Faculty (AUJEF)*, 6(3), 294- 311.
- Mueller, J. (2005). The authentic assessment toolbox: enhancing student learning through online faculty development. *Journal of Online Learning and Teaching*, 1(1), 1-7.
- Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). SAGE Publications.
- Neittaanmäki, R., & Lamprianou, I. (2024). Communal factors in rater severity and consistency over time in high-stakes oral assessment. *Language Testing*. Advance online publication. <https://doi.org/10.1177/02655322241239363>
- Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods*, 16(1), 1–13.
- Nunan, D. (1991). *Language Teaching Methodology: A Textbook for Teachers*. Prentice Hall.
- Ockey, G. J., & Li, Z. (2015). New and not so new methods for assessing oral communication. *Language Value*, 7, 1-21.
- O’Grady, S. (2019). The impact of pre-task planning on speaking test performance for English-medium university admission. *Language Testing*, 36(4), 505-526.
- O’Sullivan, B. (2000). Exploring gender and oral proficiency interview performance. *System*, 28(3), 373–386. [https://doi.org/10.1016/S0346-251X\(00\)00017-8](https://doi.org/10.1016/S0346-251X(00)00017-8)
- Oxford, R. L. (1990). *Language Learning Strategies: What Every Teacher Should Know*. New ISSN 2311-6080 York: Newbury House Publishers.
- Pallant, J. (2020). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS* (7th ed.). McGraw-Hill Education.
- Qasemi, A. S. (2020). Factors affecting students’ speaking performance in English department of Jawzjan University. *International Journal of Innovative Research and Scientific Studies*, 3(4), 129-134.
- Richards, J. C. (2008). *Teaching Listening and Speaking: From Theory to Practice*. Cambridge University Press.
- Richards, J. C., & Rodgers, T. S. (2014). *Approaches and methods in language teaching* (3rd ed.). New York, NY: Cambridge University Press.
- Ritchie, J. e., & Lewis, J. 1. e. (2003). *Qualitative research practice: A guide for social science students and researchers*. Sage Publications.
- Rost, M. (2016). *Teaching and researching listening* (3rd ed.). London, UK: Routledge.
- Saka, İ, & Merç, A. (2021). The relationship among Turkish EFL learners’ willingness to communicate in English, self-efficacy perceptions and linguistic self-confidence. *The Literacy Trek*, 7(2), 1-36. <https://doi.org/10.47216/literacytrek.770371>

- Salaberry, M. R., & Burch, A. R. (Eds.). (2021). *Assessing speaking in context: Expanding the construct and its applications*. Multilingual Matters.
- Saldaña, J. (2011). *The Coding Manual for Qualitative Researchers*. Sage Publications.
- Sato & S. Loewen (Eds.), *Evidence-based second language pedagogy: A collection of instructed second language acquisition studies* (pp. 263–286). Routledge.
- Seligman, M. E. P., & Csikszentmihalyi, M. (2000). Positive psychology: An introduction. *American Psychologist*, 55(1), 5–14. <https://doi.org/10.1037/0003-066X.55.1.5>
- Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Pearson Education.
- Shumin, K. (2002). *Factors to Consider: Developing Adult EFL Students Speaking Abilities*. In J. C. Richards, & W. A. Renandya (Eds.), *Methodology in Language Teaching* (pp. 204-211). Cambridge: Cambridge University Press.
- Souri, D., & Merç, A. (2019). Facilitating and improving speaking accuracy through self-correction and self-awareness in TOEFL independent speaking tasks. *Journal of Education, Innovation and Communication*, 1(2), 21-36.
- Swain, M. (2005). The output hypothesis: Theory and research. In *Handbook of research in second language teaching and learning* (pp. 471-483). Routledge.
- Şahin, M. D. & Öztürk, G. (2019). Mixed method research: Theoretical foundations, designs and its use in educational research. *International Journal of Contemporary Educational Research*, 6(2), 301-310. DOI: <https://doi.org/10.33200/ijcer.574002>
- Tashakkori, A., & Teddlie, C. (2003). *Handbook of Mixed Methods in Social & Behavioral Research*. Sage Publications.
- Tavakoli, P. (2009). Investigating task difficulty: Learners' and teachers' perceptions. *International Journal of Applied Linguistics*, 19(1), 1-25.
- Tavakoli, P., & Skehan, P. (2020). Strategic planning, task structure, and performance testing. In P. Skehan (Ed.), *Task-Based Language Learning and Teaching: Theoretical, Methodological, and Pedagogical Perspectives* (pp. 223-245). Routledge.
- Thornbury, S., & Slade, D. (2006). *Conversation: From description to pedagogy*. Cambridge, UK: Cambridge University Press.
- Tong, T. T. N. (2022). Factors affecting English major students' speaking performance at University of Phan Thiet. *American Journal of Sciences and Engineering Research*, 5(4), 48.
- Tsang, W. K., & Chow, K. W. (2011). The impact of novice and experienced teachers on student motivation and learning outcomes. *Educational Research*, 53(3), 283-303.
- Tuan, N. H. & Mai, T. N. (2015). Factors affecting students' speaking performance at Le Thanh Hien high school. *Asian Journal of Educational Research Vol*, 3(2), 8-23.
- Ur, P. (1996). *A course in language teaching*. Cambridge: Cambridge University Press.
- Üstün, S. (2010). Assessment Practices in EFL Contexts: Implementation of Communicative Language Teaching Principles. *Turkish Journal of Educational Studies*, 2(1), 47-62.
- Weir, C. J. (1993). *Understanding and developing language tests*. New York: Prentice Hall.

- Weir, C. J. (2005). *Language Testing and Validation: An Evidence-Based Approach*. Palgrave Macmillan.
- Wiggins, G. (1990). The case for authentic assessment. *Practical Assessment, Research, and Evaluation*, 2(2).
- Winke, P., Gass, S., & Myford, C. (2012). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, 30(2), 231–252. <https://doi.org/10.1177/0265532212456968>
- Woodrow, L. (2006). Anxiety and speaking English as a second language. *RELC Journal*, 37(3), 308–328.
- Yılmaz, F. (2015). The Impact of Test Anxiety on Turkish EFL Learners' Speaking Performance. *Journal of Language and Linguistic Studies*, 11(2), 183-198.
- Young, D. J. (1991). *Creating a low-anxiety classroom environment: What does language anxiety research suggest?* *The Modern Language Journal*, 75(4), 426–439. <https://doi.org/10.1111/j.1540-4781.1991.tb05378.x>
- Yule, G. (2010). *The study of language* 4th edition. Cambridge.
- Zechner, K., & Evanini, K. (2016). Automated speech scoring. In D. Tsagari & J. Banerjee (Eds.), *The Routledge handbook of second language assessment* (pp. 219–232). Routledge.

APPENDICES

Appendix A – Instructor Questionnaire

Dear participant,

This questionnaire is prepared for the doctoral thesis study of the Department of English Language Teaching, Faculty of Education at Anadolu University.

The purpose of this questionnaire is **to determine the major factors that have an impact on EFL learners' oral performances in the achievement tests (specifically midterm speaking and final speaking tests) at AUSFL.**

It will take approximately 10 minutes to answer the questions in the questionnaire. If you have further questions, please do not hesitate to contact me via merica@anadolu.edu.tr.

Participation in the survey is entirely voluntary. Your answers will only be used for scientific purposes and will not be shared with other individuals or institutions. Your participation in this study is highly valuable.

Thank you for your invaluable support.

Instructor
Meriç Akkaya Önal

PART I: Demographic Information

1. Gender

- ☐ Female
☐ Male
☐ NS

2. Age

- ☐ 20 - 30 ☐ 51 - 60
☐ 31 - 40 ☐ 61 - 67
☐ 41 - 50

3. Educational Background

- ☐ BA
☐ MA
☐ PhD

4. Teaching Experience

- ☐ 1 - 5 years
☐ 6 - 10 years
☐ 11 - 15 years
☐ More than 15 years

PART II: Factors Affecting Students' Speaking Performance in Their Achievement Tests (FASSP)

(3 = much, 2 = some, 1 = little, 0 = none)

	3	2	1	0
A. Rater-based factors				
In your opinion, to what extent do the following factors affect EFL learners' speaking scores in their oral achievement tests?				
<i>Speaking exam raters' ...</i>				
1. rater cognition (conscious/unconscious mental processes experienced during scoring such as observing, interpreting, inferencing, making decisions, etc.)				
2. rater characteristics (e.g. outgoing, friendly, etc.)				
3. rater expertise (experience in testing speaking)				
4. rater's support (e.g. positive body language, supportive mimes and gestures, creating positive atmosphere, etc.)				
5. following the criteria/rubric completely				
6. use of warm-up / ice-breaking strategies before the exam starts				
B. Examinee-based factors				
In your opinion, to what extent do the following factors affect EFL learners' speaking scores in their oral achievement tests?				
a. Affective factors				
<i>EFL learners'</i>				
7. motivation to speak				
8. self-confidence				
9. willingness to communicate				
10. feeling embarrassed				
11. fear of making mistakes				
12. fear of not being fully understood by the rater(s)				
13. fear of not being fully understood by the partner(s) in the speaking exam (if there is one)				
b. Linguistic factors				
<i>EFL learners' ...</i>				
14. use of communicative strategies (e.g. compensation, paraphrasing, reduction, etc.)				
15. vocabulary range				
16. appropriate use of vocabulary				
17. grammar / language use (range - use of various grammatical structures)				
18. grammar / language use (accuracy - using grammatical structures accurately)				
19. pronunciation				
20. fluency				
c. Other examinee-related factors				
<i>EFL learners' ...</i>				
21. listening ability				

22. partner in the exam (e.g. being paired with a weak/strong/dominant student, if there is one.)				
23. active participation in speaking activities in the classroom				
24. commitment outside the classroom (e.g., joining speaking clubs, watching English TV series, etc.)				
25. order of performance during the exam (speaking first or second when the exam is taken in pairs)				
26. preparation for the speaking exam				
C. Task-based Factors In your opinion, to what extent do the following factors affect EFL learners' speaking scores in their oral achievement tests?				
27. Task type (e.g. interview, role-play, description, etc.)				
28. Number of tasks (having more than one task)				
29. Variety of tasks (having various tasks)				
30. Topic knowledge (familiarity with the topic)				
31. Task quality (e.g. validity, reliability, practicality, etc.)				
32. Task difficulty				
D. Criteria/Rubric-based Factors In your opinion, to what extent do the following factors affect EFL learners' speaking scores in their oral achievement tests?				
33. Criteria/Rubric used for rating learners' performance				
34. Students' familiarity with the rubric				
E. Administration Conditions In your opinion, to what extent do the following factors affect EFL learners' speaking scores in their oral achievement tests?				
35. Physical conditions of the exam room (e.g. lighting, ventilation, size, etc.)				
36. Duration / Time allowed to perform the task (s) (e.g. limited time vs. ample time)				
37. Exam slot time (e.g. 1st slot, the slot after the break, last slot, etc.)				
38. Being video-recorded				

If there are any other factors you would like to specify, please indicate these possible factors below:

Appendix B – Student Questionnaire

Sayın Katılımcı,

Bu anket, Anadolu Üniversitesi Eğitim Fakültesi İngiliz Dili Eğitimi Anabilim Dalı doktora tez çalışması için hazırlanmıştır.

Bu anketin amacı, Anadolu Üniversitesi Yabancı Diller İngilizce Bölümü programındaki vize konuşma ve final konuşma sınavlarında İngilizceyi yabancı dil olarak öğrenen öğrencilerinin sözlü performansları üzerinde etkili olan ana faktörleri belirlemektir.

Anketteki soruları cevaplamak yaklaşık 10 dakika sürecektir. Başka sorularınız varsa, lütfen merica@anadolu.edu.tr adresinden benimle iletişime geçmekten çekinmeyin.

Ankete katılım tamamen gönüllülük esasına dayanmaktadır. Cevaplarınız sadece bilimsel amaçlar için kullanılacak ve başka kişi veya kurumlarla paylaşılmayacaktır. Bu çalışmaya katılımınız son derece değerlidir.

Desteğiniz için teşekkür ederim.

Öğretim Görevlisi
Meriç Akkaya Önal

BÖLÜM I. Demografik Bilgiler

1. Cinsiyet
☐ Kadın
☐ Erkek
☐ Belirtmek istemiyorum.
2. Yaş
☐ 16 – 20 ☐ 21 - 25 ☐ 25 üstü
3. Ne kadar zamandır İngilizce öğreniyorsunuz?
☐ 1 yıldan az ☐ 2 - 5 yıl ☐ 6- 10 yıl ☐ 11 - 15 yıl ☐ 15'ten fazla
4. Okuduğunuz lise hangisiydi?
☐ Sosyal Bilimler Lisesi
☐ Fen Lisesi
☐ Anadolu Lisesi
☐ Çok Programlı Lise
☐ Mesleki Anadolu ve Teknik Anadolu Lisesi
☐ İmam Hatip Lisesi
☐ Özel Okul/Kolej
☐ Diğer: _____
5. Şu an okuduğunuz dil düzeyi nedir?
☐ A+ ☐ C
☐ A ☐ D
☐ B

6. Hazırlık eğitiminizi tamamladığınızda okuyacağınız bölüm nedir?

BÖLÜM II: Öğrencilerin Konuşma Sınavlarındaki Performanslarını Etkileyen Faktörler

(3 = çok, 2 = biraz, 1 = az, 0 = hiç)

	3	2	1	0
A. Değerlendirici Temelli Faktörler Sizce aşağıdaki faktörler İngilizceyi yabancı dil olarak öğrenen öğrencilerin vize ve final konuşma sınavlarındaki notunu ne ölçüde etkiler?				
Konuşma sınavı değerlendiricisinin ...				
1. puanlayıcı bilişi (gözlemleme, yorumlama, çıkarım yapma, karar verme vb. gibi puanlama sırasında yaşanan bilinçli/bilinçsiz zihinsel süreçler)				
2. puanlayıcının karakteristik özellikleri (örn. dışa dönük, arkadaş canlısı, vb.)				
3. puanlayıcı uzmanlığı (konuşma testinde deneyimli olma)				
4. değerlendiricinin desteği (örn. olumlu beden dili, destekleyici mimik ve jestler, olumlu atmosfer yaratma, vb.)				
5. kriterleri/rubriği tamamen takip etmesi				
6. sınav başlamadan önce ısınma stratejilerinin kullanılması (örn. genel bir sosyal sohbet)				
B. Öğrenci Temelli Faktörler Sizce aşağıdaki faktörler İngilizceyi yabancı dil olarak öğrenen öğrencilerin vize ve final konuşma sınavlarındaki notunu ne ölçüde etkiler?				
a. Duygusal Faktörler İngilizceyi yabancı dil olarak öğrenen öğrencilerin...				
7. konuşma motivasyonu				
8. özgüveni				
9. iletişim kurma isteği				
10. utangaç hissetmesi				
11. hata yapma korkusu				
12. değerlendirici(ler) tarafından tamamen anlaşılamama korkusu				
13. varsa sınav partneriniz/partnerleriniz tarafından tamamen anlaşılamama korkusu				
b. Dilsel Faktörler İngilizceyi yabancı dil olarak öğrenen öğrencilerin...				
14. iletişim stratejileri kullanımı (örn. telafi etme, başka şekilde açıklamaya çalışma, vb.)				
15. kelime dağarcığı aralığı				
16. uygun kelime kullanımı				
17. dilbilgisi / dil kullanımı (çeşitli dilbilgisi yapılarının kullanımı)				
18. gramer / dil kullanımı (doğru dilbilgisi yapılarının kullanımı)				
19. telaffuz				
20. akıcılık				

c. Öğrenci temelli diğer faktörler				
İngilizceyi yabancı dil olarak öğrenen öğrencilerin...				
21. dinleme yeteneği				
22. varsa sınavdaki partner(ler)i (örn. zayıf/güçlü/baskın bir öğrenciyle eşleştirilmek)				
23. sınıf içi konuşma aktivitelerine aktif katılımı				
24. sınıf dışında da İngilizceye yönelik çabaları(örn. konuşma kulüplerine katılmak, İngilizce TV dizileri izlemek vb.)				
25. sınav sırasındaki performans sırası (sınav çiftler halinde yapıldığında ilk veya ikinci konuşmacı olma durumu)				
26. konuşma sınavına hazırlanması				
C. Sınav Sorusuna Dayalı Faktörler				
Sizce aşağıdaki faktörler İngilizceyi yabancı dil olarak öğrenen öğrencilerin vize ve final konuşma sınavlarındaki notunu ne ölçüde etkiler?				
27. Sınavda sorulan sorunun türü (örn. mülakat, soru-cevap, rol yapma, vb.)				
28. Soru sayısı (birden fazla soru olması)				
29. Soru çeşitliliği (farklı türlerde soruların olması)				
30. Konu bilgisi (konuya aşinalık)				
31. Soru kalitesi (örn. geçerlilik, güvenilirlik, pratiklik, vb.)				
32. Sorunun zorluğu				
D. Kriter/Rubrik Temelli Faktörler				
Sizce aşağıdaki faktörler İngilizceyi yabancı dil olarak öğrenen öğrencilerin vize ve final konuşma sınavlarındaki notunu ne ölçüde etkiler?				
33. Öğrencinin performansını ölçmede kullanılan kriter/rubrik				
34. Öğrencinin kullanılan kritere/rubriğe olan aşinalığı				
E. Fiziksel Koşullar				
Sizce aşağıdaki faktörler İngilizceyi yabancı dil olarak öğrenen öğrencilerin vize ve final konuşma sınavlarındaki notunu ne ölçüde etkiler?				
35. Dersliğin fiziksel kalitesi (örn. aydınlatma, havalandırma, sınıfın boyutu, vb.)				
36. Sınavı gerçekleştirmek için izin verilen süre / Zaman (örn. zamanın gereğinden fazla ya da az olması)				
37. Sınav oturum zamanı (örn. 1. oturum, ara verildikten sonraki oturum, son oturum gibi)				
38. Video kayıt altına alınmak				

Belirtmek istediğiniz başka faktörler var ise, lütfen aşağıdaki boşluğa yazınız.

Appendix C – Extracts from Expert Opinions on the Questionnaire Items

rater cognition (mental processes experienced during rating process)	Comments:
Appropriate for use	
Appropriate for use	
Appropriate for use	
Appropriate for use after necessary changes	such as fatigue, sickness can be added to make it more to point
Appropriate for use	
Appropriate for use after necessary changes	Here it may be a good idea to give an example or an indicator of what we mean by mental processes, because everyone may not be thinking about the same factors while answering this item and this may affect the results.

fear of not being understood by the partner	Comments:
Appropriate for use after necessary changes	"Partner in the exam" or "pair in the exam" I think 'partner' doesn't sound just right. I still understand what you are referring to, but something feels off for me :)
Appropriate for use	
Appropriate for use after necessary changes	This item actually depends on how the speaking test is designed/administered. For instance, for our case in our school, this item is suitable; however, if the student takes the exam alone, then there will be no partner of that student during the exam. If we think that your study will be published/presented in an international setting, where examinees are admitted to the speaking exam alone, then the audience may not understand what you mean with the word "partner"
Appropriate for use after necessary changes	In some exam sessions, there is not any partner, so this item might be named as "fear of not being understood by the partner (if available)".
Appropriate for use after necessary changes	if the participant students are familiar with the speaking exams with the partners, no problem with this item. But if the speaking exam is teacher-student format, then they may not understand the item easily. In that case, in student-student format can be added..But just an idea...
Appropriate for use	
Appropriate for use after necessary changes	Same as above

Appendix D – Extracts from Expert Opinions on the Semi-structured Interview Questions

2. What can be the factors that affect your students' speaking performances in Midterm and Final exams? Why?	Comments
Appropriate for use	
Appropriate for use	
Appropriate for use	
Appropriate for use	
Appropriate for use	
Appropriate for use	
Appropriate for use after necessary changes	Since this is a very broad question, it may be difficult to categorize the answers, maybe the factors can be asked in sub-headings.

[illegible]

Appendix E – Semi-structured Interview Questions

Semi-structured interview questions for instructors/test designers/administrators

1. Are you content/satisfied with the speaking grades of your students? Why? Why not?
2. What can be the factors that affect your students' speaking performances in Midterm and Final exams? Why?
3. What do you think is the biggest challenge for your students in the speaking exams, specifically midterm and final exams?
4. What makes it easier for your students in the speaking exams, specifically midterm and final exams?
5. In your questionnaire results, ----- is found to be the most important factor. Why do you think it is the most important factor?
6. In your questionnaire results, ----- is found to be the least important factor. Why do you think it is the least important factor?
7. What can be done to help your students demonstrate higher achievement (as an instructor/ test designer/ a member of school administration)?

Semi-structured interview questions for students

1. Are you content/satisfied with the speaking grade you got in the midterm exam? Why? Why not?
2. What can be the factors that affect your speaking performances in Midterm and Final exams? Why?
3. Think of your latest speaking exam. Why do you think you got this score?
4. What factors affected you the most in the latest speaking exam?
5. What do you think is the biggest challenge for you in the speaking exams, specifically midterm and final exams?
6. What makes it easier for you in the speaking exams, specifically midterm and final exams?
7. In your questionnaire results, ----- is found to be the most important factor. Why do you think it is the most important factor?
8. In your questionnaire results, ----- is found to be the least important factor. Why do you think it is the least important factor?
9. What can you do to demonstrate higher achievement in the speaking exams?

Appendix F – Extracts from Expert Opinions on the Video Stimulated Recall Interview Questions

3. What factors were at play when you gave this score?	Comments:
	What factors do you think affected you when you gave this score? I feel like we might want to make the question clearer?
Appropriate for use	
Appropriate for use after necessary changes	"this score", what do you refer to with 'this'? I mean which score? the fair or unfair score in the previous question? or the score in the second grading?
Appropriate for use	
Appropriate for use	
Appropriate for use	
Appropriate for use after necessary changes	In this part, instead of asking which factors, it seems to me that if we ask him to justify the grade he gave with the reasons in detail , maybe he can also explain it by directing it to the criteria as well.

[illegible]

Appendix G – Video-stimulated Recall Interview Questions

Video stimulated recall interview questions for instructors

1. Please grade the learner again.
If the score is same/similar, why do you think you gave a same/similar score
If the score is different, why do you think it is different now?
2. Do you think your first score is fair enough? Why? Why not?
3. What factors were at play when you gave this score?
4. Would you change anything in this scoring session? E.g. the task, student's partner, etc.?

Video stimulated recall interview questions for students

1. Do you find your grade fair enough? Why? Why not?
2. You lost some points here. Why do you think you lost points here?
3. You said ... here. What made you say so?
4. What factors affected you during the exam? Why?
5. If you had the chance to take the exam again, what factors would you take into consideration to get a better score? / what would you do differently? Why?

Appendix H – Sample Color-code Labeling from the interviews

Cevap 17: Birinci sebep sanırım **öğrencilerin kaygı seviyesi çok yüksek** konuşma sınavında aşırı heyecan yaptıklarını söylediler bilmelerine rağmen o an algılayamadıklarını kilitlendiğini çok heyecan yaptıklarını söylediler. 1-2 tanesi bazı **kelimelere takılmış ve kelimeyi anlayamadığı için** de ne kadar soru tekrar edilirse edilsin, o keywordi çözemediği için cevaplayamadığını söyledi. Başka bir sebep, **belki öğrenciler hani o konunun konuşulduğu derse katılmadı ve konuya dair çok aklına fikir gelmedi** o an öğrencilerimiz geriye dönüp çok çalışmıyorlar biliyorsun. Geri dönüp hani ben bu hafta katılmamıştım, ne konuşmuşuz, derste demiyorlar, normal devam ediyorlar. O haftayı boşlukta bırakıp dolayısıyla konuya hakim olmayıp **fikir üretmediklerini** düşünüyorum ya da **keywordü** hatırlayamayıp kilitlendi ya da sorudaki **keywordü** anlayamadılar. Dolayısıyla da o soruya cevap veremediler. Diğer düzeylerde de yine arkadaşlarımızın hani anketlerden ya da onlarla konuştuğumda aldığım izlenim onların da bir memnuniyetsizlik içerisinde olması. Öğrencilerin genel olarak **çok fazla fikir üretmediklerine** dair bir bilgin var ama çok da detaylı bilgin yok, onun için yanlış bilgi de vermek istemem.

Cevap 14: Birincisi en başta **öğrencilerin hazır bulunma düzeyleri ya da istekleri**. Genelde Türk öğrenciler konuşma konusunda biraz isteksiz olurlar. Bunun da çeşitli nedenleri var. Muhtemelen bizim ortaokulda, ilkokulda, ortaokulda ve lisede konuşma derslerine yönelik çok fazla bir etkinlik olmadığı için öğrenciler daha pasif kalıyorlar. Dolayısıyla hazırlık sınıfına geldikleri zaman da konuşma dersini, hazırlayan bahsetmişim, öğretmen konuşacağı öğrencinin dinleyeceği bir dersmiş gibi. Pasif kalacağı bir dersmiş gibi görebiliyorlar. İkincisi **bazı öğrenciler derslerde ne kadar katılım istesek de çok fazla katılımcı olmayabiliyorlar**. Onları yönlendirmeye çalışsak da bir kısımda eğer belli bir konuda ya da genelde kendini kapatıyorsa ona yapabilecek çok fazla bir şey olmuyor. **Diğeri sınav sırasında sınav günü ya da sınav sırasındaki faktörler** etkili olabilir. **Arkadaşlarının tavırları etkili** olabilir çünkü bu öğrenciler konuşma sınavından çıktıktan sonra bir önceki öğrenci grubunu pairini dinliyorlar. O işte onlardan bilgi alıyorlar. Ne çıktı? Hocalar iyi miydi işte anlayışlar mıydı? Filan gibi onlar etkili olabilir. **Soru etkili olabilir** en önemlisi. Bazen sorular sorun olabiliyor. Tek bir soru soruluyor, o soruyu sorulduktan sonra bir kez daha aynı soru aynı şekilde tekrar ediliyor. Bir üçüncü şans bazen veriliyor. Genelde **paraphrasing** yapmaya ya da **synonym** kullanmaya izin verilmediği için öğrenci o cümle o soru kalıbının içerisindeki herhangi bir kelimeyi, ya da yapıyı anlamadığında soruyu yanlış anlayabiliyor, hiç anlaya anlamayabiliyor. Bu nedenle beklenen performansı gösteremeyebiliyorlar. Bir diğeri **hocaların tavırları önemli olabilir**. Burada hocaların daha anlayışlı olması çeşitli ipuçları vererek feedback vermesi sınav sırasında tamam kopya vermiyoruz ama en azından bir jestlerle, mimiklerle konuşmaya katıldığınızı aktif dinlemeye olduğumuzu gösteren işaretler olmalı. Bizim galiba en büyük sıkıntımız konuşmayı monolog bir yapıda algılamamızdan kaynaklanıyor. Oysa konuşma dediğimiz şey hatta dinleme bile karşılıklı olan bir etkileşim. Bu etkileşimi sağlayan bir ortamın olması gerekir diye düşünüyorum. Bir diğer sebep **kamera kaydı almıyor** bizim okulumuzda. Kameradan dolayı öğrenciler belki endişelenmiş olabilirler.

U **user**
Yüksek kaygı
Kelimeleri anlamama
Tekrar yapmamları
Fikir üretememe

U **user**
Öğrencileri hazır bulunma düzeyleri
Katılım isteği
Sınav sırasında sınav günü ya da sınav sırasındaki faktörler → arkadaşlarının sorular hakkındaki yorumları
Soru → bazen tek soru olması ya da paraphrase yapılamaması
Hoca tavrı
Kamera kaydı

Appendix I - Interview results of instructors (n=22) and students (n=15)

	Factors	Instr.	St.	Challenges	Instr.	St.	Facilitators	Instr.	St.
Examinee-based	Affective factors - anxiety	16	7	Affective factors - anxiety	17	8	Achievement-oriented exam	12	
	Readiness - practice	16	9	Readiness - practice	8		Practice	8	10
	Linguistic factors – lg prof.	10		Linguistic factors – lack of voc.	6	5			
	Content production – limited content production	5		Content production – Inability to generate ideas	10				
Rater-based	Behavior	5	4	Behavior	8	1	Friendly behavior	4	5
	Familiarity		3	Familiarity	5	1	Familiar instructors	7	5
Contextual	Task questions	4	3	Exam-related context - Task questions	5	4	Exam questions	4	4
						(Difficulty understanding due to remote or unclear format + unprovided written questions + new topic)			(familiar topics + visuals)
				Physical context – video-recording	6	1			
				Peer effect - unfamiliar partners	5	2	Peer effect - Exam with classmates (familiar partner)	6	3
				Peer effect - Negative comments influence	-	2	Peer effect - Negative comments influence		1

Appendix J - Overall VSR interview results of instructors (n=9) and students (n=15)

Factors	Instructors' viewpoint	Students' viewpoint
Examinee-based	Affective factors - anxiety	2
	Affective factors (other)	1
	Self-perception	6
	Readiness	5
	Ling. factors	4
Rater-based	Behavior	4
	Familiarity	3
Context-based	Peer effect - familiar partner	2
	Peer effect (other)	4
	Exam conditions	3
	Exam environment	5
	Criteria	3
	Exam questions	2
		-
		1

Appendix K – Ethics Committee Declaration (Etik Kurul Beyannamesi)

Evrak Kayıt Tarihi: 12.09.2023

Protokol No: 593455

Tarih: 03.10.2023



ANADOLU ÜNİVERSİTESİ
SOSYAL VE BEŞERÎ BİLİMLER BİLİMSEL ARAŞTIRMA VE YAYIN ETİĞİ KURULU
KARAR BELGESİ

ÇALIŞMANIN TÜRÜ:	Doktora Tez Çalışması
KONU:	Eğitim Bilimleri
BAŞLIK:	İngilizceyi Yabancı Dil Olarak Öğrenen Üniversite Öğrencilerinin Konuşma Performanslarını Etkileyen Faktörler
PROJE/TEZ YÜRÜTÜCÜSÜ:	Prof. Dr. Özgür YILDIRIM
TEZ YAZARI:	Meriç AKKAYA ÖNAL
ALT KOMİSYON GÖRÜŞÜ:	-
KARAR:	Olumlu

RESUME

Name: Meriç AKKAYA ÖNAL

Education

Anadolu University – Faculty of Education
BA in English Language Teaching
2000 – 2004

İhsan Doğramacı Bilkent University – Graduate School
MA in MATEFL
2013 – 2015

Anadolu University –Graduate School
Ph.D. Program in English Language Teaching
2023 – 2025

Teaching Experience

Mehmet Akif Primary School, Kırıkhan Hatay
2004 – May 2006

Anadolu University School of Foreign Languages, Eskişehir
2006 – August 2023

Anadolu University Faculty of Education Department of English Language Teaching, Eskişehir
2023 – continues

Skills

Microsoft Office & Canva

Languages

English (C1), German (A2)

Certificates

CELTA – IH London - 2013

DELTA – ITI (International Training Institute) - 2019

References

Available upon request.