



**TÜRK İNGİLİZCE OKUTMANLARININ YABANCI
DİLDE ÖLÇME DEĞERLENDİRME OKURYAZARLIĞI:
ÇOKLU BİR DURUM ÇALIŞMASI**

Ahmet Erdost YASTIBAŞ

**Doktora Tezi
Yabancı Diller Eğitimi Ana Bilim Dalı
Prof. Dr. Mehmet TAKKAÇ
2018
(Her Hakkı Saklıdır)**

T.C.
ATATÜRK ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ
YABANCI DİLLER EĞİTİMİ ANA BİLİM DALI
İNGİLİZCE EĞİTİMİ BİLİM DALI

TÜRK İNGİLİZCE OKUTMANLARININ YABANCI DİLDE ÖLÇME
DEĞERLENDİRME OKURYAZARLIĞI: ÇOKLU BİR DURUM
ÇALIŞMASI

(Language Assessment Literacy of Turkish EFL Instructors: A Multiple-case
Study)

DOKTORA TEZİ

Ahmet Erdost YASTIBAŞ

Danışman: Prof. Dr. Mehmet TAKKAÇ

ERZURUM

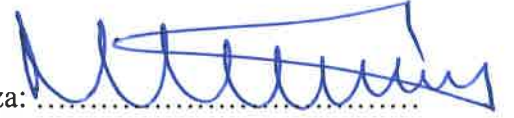
Ocak, 2018

KABUL VE ONAY

Prof. Dr. Mehmet TAKKAÇ danışmanlığında, Ahmet Erdost YASTIBAŞ tarafından hazırlanan “Türk İngilizce Okutmanlarının Yabancı Dilde Ölçme Değerlendirme Okuryazarlığı: Çoklu Bir Durum Çalışması” başlıklı çalışma 19/01/2018 tarihinde yapılan savunma sınavı sonucunda başarılı bulunarak jürimiz tarafından Yabancı Diller Eğitimi Anabilim Dalı’nda Doktora Tezi olarak kabul edilmiştir.

Başkan/Danışman: Prof. Dr. Mehmet TAKKAÇ

İmza:



Jüri Üyesi : Doç. Dr. Mustafa Zeki ÇIRAKLI

İmza:



Jüri Üyesi : Yrd. Doç. Dr. Ali DİNÇER

İmza:



Jüri Üyesi : Yrd. Doç. Dr. Muzaffer BARIN

İmza:



Jüri Üyesi : Yrd. Doç. Dr. Oktay YAĞIZ

İmza:



Yukarıdaki imzaların adı geçen öğretim üyelerine ait olduğunu onaylarım.

01 Subat 2018 / 2018



Prof. Dr. Mustafa SÖZBİLİR

Enstitü Müdürü

TEZ ETİK VE BİLDİRİM SAYFASI

Doktora Tezi olarak sunduđum “TÜRK İNGİLİZCE OKUTMANLARININ YABANCI DİLDE ÖLÇME DEĞERLENDİRME OKURYAZARLIĐI: ÇOKLU BİR DURUM ÇALIŞMASI” başlıklı çalışmanın, bilimsel ahlak ve geleneklere aykırı düşecek bir yardıma başvurmaksızın yazıldığını ve yararlandığım eserlerin kaynakçada gösterilenler olduğunu, bunlara atıf yapılarak yararlanılmış olduğunu belirtir ve onurumla doğrularım.

Lisansüstü Eğitim-Öğretim yönetmeliğinin ilgili maddeleri uyarınca gereğinin yapılmasını arz ederim.

- Tezimin tamamını her yerden erişime açılabilir.
- Tezim sadece Atatürk Üniversitesi yerleşkelerinde erişime açılabilir.
- Tezimin **1.** yıl süreyle erişime açılmasını istemiyorum. Bu sürenin sonunda uzatma için başvurmadığım takdirde, tezimin tamamını her yerden erişime açılabilir.

01/02/2018

Ahmet Erdost YASTIBAŞ

ayalın

ÖZET

DOKTORA TEZİ

TÜRK İNGİLİZCE OKUTMANLARININ YABANCI DİLDE ÖLÇME DEĞERLENDİRME OKURYAZARLIĞI: ÇOKLU BİR DURUM ÇALIŞMASI

Ahmet Erdost YASTIBAŞ

2018, 245 sayfa

Ölçme değerlendirme alanıyla ilgili yapılan çalışmalarda yabancı dilde ölçme değerlendirme okuryazarlığı giderek artan bir şekilde önemli hale gelmiştir; ancak yapılan çalışmalar, ölçme değerlendirme okuryazarlığının yabancı dil öğretmenleri tarafından sınıflarında nasıl uygulandığına çok fazla odaklanmamışlardır. Bu yüzden bu çalışma, yabancı dilde ölçme değerlendirme okuryazarlığının sınıfta nasıl uygulandığını açıklamayı amaçlamaktadır. Çalışma, bir Türk üniversitesinde çalışan sekiz Türk İngilizce öğretmeni ile gerçekleştirilmiştir. Araştırma deseni olarak çoklu durum çalışması kullanılmıştır. Veriler; bireysel görüşmeler, sesli düşünme, gözlemler, odak grup çalışması ve doküman analizi ile toplanmıştır. Çalışmanın inandırıcılığı arttırmak için üçgenleme, yoğun anlatım gibi yöntemler kullanılmıştır. Toplanan veriler, içerik analizi kullanılarak analiz edilmiştir. Bulgulara göre katılımcıların, ölçme değerlendirme faaliyetlerinde eleştirel bir tutum sergiledikleri görülmüştür. Eğitim fakültelerinden mezun olan katılımcıların ölçme değerlendirme faaliyetlerinde aldıkları ölçme değerlendirme dersinin etkili olduğu görülürken farklı fakültelerden mezun olarak formasyon alan katılımcılarda bu tür eğitimlerin etkili olmadığı görülmüştür. Katılımcıların ölçme değerlendirmede kendilerini deneyimleyerek geliştirdikleri tespit edilmiştir. Katılımcıların geçerlik, güvenilirlik ve ölçmede hata gibi temel kavramlarla ilgili olarak kendi tanımlarını geliştirdikleri ve bu tanımlara göre çeşitli teknikler kullanarak sınavlarını geçerli ve güvenilir yapmaya çalıştıkları görülmüştür. Katılımcıların; geçerlik anlayışlarının kullandıkları ders kitaplarıyla sıkı bir şekilde bağlantılı olduğu, kapsam geçerliğini ön planda tuttukları ve kapsam geçerliliğinin güvenilirlik anlayışlarıyla sıkı sıkıya bağlı olduğu bulunmuştur. Öğrenci sayısının fazlalığı, iş yüklerinin ağır olması gibi durumların ölçme değerlendirme

okuryazarlığının yedi alt yeterliliğini etkilediği bulunmuştur. Ayrıca bazı katılımcıların, sahip oldukları deneyime ve eğitim – öğretim anlayışına göre ölçme değerlendirme inisiyatif olarak diğer katılımcılarından farklılaştığı görülmüştür. Çalışmanın son kısmında çalışmada elde edilen sonuçların, yabancı dilde ölçme değerlendirme eğitimlerinin geliştirilmesindeki muhtemel katkılarına değinilmiştir.

Anahtar Kelimeler: Yabancı dilde ölçme değerlendirme okuryazarlığı, Eleştirel yaklaşım, Kişisel gelişim



ABSTRACT

Ph. D. DISSERTATION

LANGUAGE ASSESSMENT LITERACY OF TURKISH EFL INSTRUCTORS: A MULTIPLE-CASE STUDY

Ahmet Erdost YASTIBAŞ

2018, 245 pages

Language assessment literacy has become increasingly important in the studies made on assessment and evaluation; however, those studies have not focused on how language assessment literacy is implemented by language teachers in their language classes. Therefore, the present study aimed at explaining how language assessment literacy is implemented in language classes. The study was made with eight Turkish instructors working at a Turkish university and teaching English as a foreign language (EFL). A multiple-case study research design was used in the study. Data were collected with individual interviews, think-aloud protocol, observations, focus group discussion and document analysis. To increase the trustworthiness of the study, several techniques including triangulation and thick description were used. The collected data were content-analyzed. According to the findings of the study, the participants were found to have a critical attitude toward assessment and evaluation. It was understood that pre-service assessment training was effective in the assessment and evaluation practices of the participants graduating from faculty of education, while such training was ineffective in the assessment and evaluation practices of the others graduating from different faculties. The participants were found to have improved themselves in language assessment and evaluation by gaining experience. The findings have also indicated that the participants developed their own definitions of the basic assessment concepts like validity, reliability and measurement error and tried to make their exams valid and reliable by using the techniques depending on their definitions. In addition, the findings have shown that the participants' understanding of validity was closely related to their course books, they paid more attention to content validity and their understanding of reliability was closely related to content validity. The factors like the

number of the students and workload were revealed to affect the seven sub-components of language assessment literacy. Besides, some participants were found to differentiate from the others as they took initiative depending on their experience and teaching approach. In the last part of the study, how the results could contribute to the development of language assessment training was mentioned.

Key Words: Language assessment literacy, Critical approach, Self-improvement



ACKNOWLEDGEMENTS

I would like to express my sincerest appreciation to my supervisor Prof. Dr. Mehmet TAKKAÇ for his academic supervision during my dissertation without which this study can not have gone forward. I can not express my deepest gratitude to him in my limited words.

In addition, I would like to thank my dissertation committee members Assist. Prof. Dr. Oktay YAĞIZ and Assist. Prof. Dr. Muzaffer BARIN. They always encouraged me and gave me invaluable comments during this study.

I am also in debt to my defense jury members Assist. Prof. Dr. Ali DİNÇER and Assoc. Prof. Dr. Mustafa Zeki ÇIRAKLI who carefully reviewed every page of this dissertation and commented their concerns generously.

Besides, I would like to thank TÜBİTAK (The Scientific and Technological Research Council of Turkey) as TÜBİTAK always supported me during my Ph. D. studies with the scholarship program of 2211-A (Graduate Scholarship Program).

Lastly, I must express my very profound gratitude to my wife Gülşah ÇINAR YASTIBAŞ, my sisters, my parents and my parents-in-law for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of searching and writing this dissertation.

Ocak-2018

Ahmet Erdost YASTIBAŞ

DEDICATION

I dedicated this dissertation to my wife, Gülşah, the most important person in my life. We encountered lots of difficulties during my dissertation studies, but we overcame them together.



CONTENT

KABUL VE ONAY	i
TEZ ETİK VE BİLDİRİM SAYFASI	ii
ÖZET	iii
ABSTRACT	v
ACKNOWLEDGEMENTS	vii
DEDICATION	viii
CONTENT	ix
LIST OF TABLES	xvi
LIST OF FIGURES	xvii
ABBREVIATIONS	xviii

CHAPTER ONE

1. INTRODUCTION	1
1.1. Introduction	1
1.2. Background of the Study	1
1.3. Statement of the Problem	3
1.4. Purpose of the Study	4
1.5. Research Questions	4
1.6. Significance of the Study	5
1.7. Definitions of the Key Terms	6
1.8. Overview of the Dissertation	6
1.9. Conclusion	7

CHAPTER TWO

2. LITERATURE REVIEW	8
2.1. Introduction	8
2.2. Theoretical Framework: Assessment for Learning	8
2.3. Assessment Literacy	10
2.3.1. Why is assessment literacy important?	11
2.4. Standards of Assessment Literacy	15
2.4.1. Choosing appropriate assessment methods for instructional purposes	16

2.4.2. Developing appropriate assessments for instructional purposes	17
2.4.3. Administering exams, scoring them and interpreting their results.....	18
2.4.4. Using assessment results in making decision.....	19
2.4.5. Developing valid grading procedures.....	20
2.4.6. Communicating assessment results	20
2.4.7. Recognizing unethical, illegal and other inappropriate assessments and uses of assessment information	21
2.5. Language Assessment Literacy.....	21
2.5.1. Why is language assessment literacy important?	23
2.5.1.1. Being the agent of language assessment	23
2.5.1.2. Language assessment training	24
2.5.1.3. Assessment culture vs. testing culture.....	25
2.5.1.4. Educational and political reforms.....	26
2.6. Language Assessment Literacy Research.....	28
2.6.3.1. Language assessment literacy research in the international sphere.....	28
2.6.3.2. Language assessment literacy research in the national sphere.....	35
2.7. Conclusion	39

CHAPTER THREE

3. METHODOLOGY	40
3.1. Introduction.....	40
3.2. Research Design.....	40
3.3. Research Setting.....	43
3.4. Participants.....	44
3.5. Data Collection Tools	46
3.5.1. Semi-structured individual interviews.....	47
3.5.2. Think-aloud Protocol.....	48
3.5.3. Focus group discussion	49
3.5.4. Semi-structured, field and non-participant observations.....	49
3.5.5. Document analysis.....	50
3.5.6. The relationships between data collection tools and sub-components of language assessment literacy	52

4.3.1. Choosing appropriate assessment methods for instructional purposes	79
4.3.1.1. Definition of measurement error	79
4.3.1.1.1. Sources of measurement error	81
4.3.1.1.2. Types of measurement errors	81
4.3.1.2. Definition of validity	81
4.3.1.2.1. Types of validity	83
4.3.1.3. Valid and invalid assessment data	83
4.3.1.3.1. The effects of valid and invalid data on the participants' instruction ...	83
4.3.1.4. The types of the assessment methods the participants used during their teaching career and for their present classes	84
4.3.1.5. Choosing assessment methods	84
4.3.1.5.1. Purposes for choosing assessment methods	87
4.3.1.5.2. Ideas about the strengths and weaknesses of the chosen assessment methods.....	87
4.3.2. Developing appropriate assessments for instructional purposes	88
4.3.2.1. Preparing exams in relation to the chosen assessment methods.....	89
4.3.2.1.1. The participants in action	90
4.3.2.1.1.1. Starting to prepare the exams	91
4.3.2.1.1.2. Choosing reading passages, listening audio and/or words to prepare the questions	92
4.3.2.1.1.3. Deciding what to ask in the exams	93
4.3.2.1.1.4. Preparing the questions.....	94
4.3.2.1.1.5. Self-assessing the written questions	97
4.3.2.1.1.6. Evaluating the available questions to use in the new exams	98
4.3.2.1.1.7. Finalizing the preparation of the exams	100
4.3.2.1.2. The participants in documents.....	101
4.3.2.1.2.1. Reading passages and questions	101
4.3.2.1.2.2. Listening audio and questions	102
4.3.2.1.2.3. Vocabulary questions	102
4.3.2.1.2.4. Grammar questions.....	103
4.3.2.1.2.5. Open-ended questions	103
4.3.2.2. The types of the questions used in the chosen assessment methods	104

4.3.2.3. Providing validity	104
4.3.3. Administering exams, scoring them and interpreting their results.....	107
4.3.3.1. Definition of reliability.....	108
4.3.3.2. Providing reliability.....	109
4.3.3.3. Administering exams.....	112
4.3.3.4. The problems encountered in administering the exams and the ways used to overcome the problems	114
4.3.3.5. Scoring.....	115
4.3.3.6. Grading.....	118
4.3.3.7. Consistency of assessment interpretation.....	119
4.3.3.8. Interpreting formal and informal student evaluation.....	120
4.3.3.9. Using student assessment results for assessment tools and students’ learning.....	121
4.3.3.10. Washback effect	121
4.3.3.11. Confidentiality of assessment and assessment results.....	122
4.3.3.12. Attitude toward exam complaint	123
4.3.4. Using assessment results in making decisions about student, instruction, school and curriculum	124
4.3.4.1. Evaluating assessment data	125
4.3.4.2. Wrong and correct interpretation of assessment data.....	125
4.3.4.3. Developing instructional plan for students.....	126
4.3.4.4. Making changes in instruction and curriculum	126
4.3.5. Developing valid grading procedures using students’ assessments	127
4.3.5.1. Grading systems	128
4.3.5.2. The reasons for using the grading systems.....	129
4.3.5.3. Developing grading systems.....	130
4.3.5.4. Validity of the grades given through the grading systems	133
4.3.5.5. Purposes for giving grades	134
4.3.6. Communicating assessment results to students and other stakeholders....	134
4.3.6.1. Communicating assessment results to students and school administration.....	135
4.3.6.2. Meanings of assessment results to different stakeholders.....	136

4.3.6.3. Correct interpretations of the results	136
4.3.6.4. Limitations in interpreting assessment data	137
4.3.6.5. Reflections of interpreting assessment data	137
4.3.6.6. Avoiding misinterpretation of assessment data.....	138
4.3.6.7. Avoiding possible measurement errors in communicating assessment results	138
4.3.7. Recognizing unethical, illegal and inappropriate assessments and uses of assessment information	139
4.3.7.1. The ethical, legal and professional assessment and evaluation practices.....	139
4.3.7.2. The problems encountered and how they were overcome	141
4.4. Conclusion	142

CHAPTER FIVE

5. DISCUSSION	143
5.1. Introduction.....	143
5.2. Implementing the Sub-components of Language Assessment Literacy in the Class	143
5.2.1. Choosing appropriate assessment methods for instructional purposes	143
5.2.2. Developing appropriate assessments for instructional purposes	146
5.2.3. Administering exams, score them and interpret their assessment results .	148
5.2.4. Using assessment results in making decisions about students, planning instruction and developing curriculum	150
5.2.5. Developing valid grading procedures which use students' assessments...	151
5.2.6. Communicating assessment results to students and school administrators	152
5.2.7. Recognize unethical, illegal and inappropriate assessments and uses of assessment information.....	153
5.3. Factors Affecting the Implementation of Language Assessment Literacy	153
5.4. The Effects of Language Assessment Literacy on the Turkish EFL Instructors ...	161
5.5. Difficulties Encountered in Implementing and Ways to Overcome Them.....	168
5.6. The Implementability of Language Assessment Literacy	172

5.7. Conclusion	173
-----------------------	-----

CHAPTER SIX

6. CONCLUSION	174
6.1. Introduction.....	174
6.2. Summary of the Study	174
6.2.1. The epistemological dimension of language assessment literacy	174
6.2.2. The ontological dimension of language assessment literacy.....	177
6.2.3. The practical dimension of language assessment literacy	178
6.3. Implications of the Study	180
6.4. Strengths of the Study	181
6.5. Limitations of the Study.....	182
6.6. Suggestions for Further Research	183
REFERENCES.....	184
APPENDICES	207
APPENDIX 1. The informed consent form for participating the study	207
APPENDIX 2. The first semi-structured interview protocol and questions	208
APPENDIX 3. The second semi-structured interview protocol and questions...	211
APPENDIX 4. The third semi-structured interview protocol and questions	215
APPENDIX 5. Think-aloud protocol	217
APPENDIX 6. The focus group discussion protocol and questions	218
APPENDIX 7. The first and second non-participant observation protocols.....	220
APPENDIX 8. The third non-participant observation protocol	221
APPENDIX 9. The fourth non-participant observation protocol.....	222
APPENDIX 10. The document analysis protocol	223
CURRICULUM VITAE.....	225

LIST OF TABLES

Table 3.1.	The Summary of Designing the Present Study in Five Stages	43
Table 3.2.	Demographic Information about Each Participant	46
Table 3.3.	The Relationship between Data Collection Tools and Sub-components of Language Assessment Literacy	52
Table 3.4.	Strategies Used to Make the Present Study Trustworthy	53
Table 4.1.	Themes and Codes of the First Interviews	60
Table 4.2.	Themes of the Second Interviews.....	78
Table 4.3.	Codes of the Focus Group Discussion.....	78
Table 4.4.	Codes of the First Sub-component LAL.....	79
Table 4.5.	Codes of the Second Sub-component of LAL.....	89
Table 4.6.	Codes of the Think-Aloud Protocols	90
Table 4.7.	Codes of the Third Sub-component of LAL.....	107
Table 4.8.	Codes of the Third Observations	107
Table 4.9.	Codes of the Fourth Observations	107
Table 4.10.	Codes of the Fourth Sub-component of LAL.....	124
Table 4.11.	Codes of the Fifth Sub-component of LAL.....	127
Table 4.12.	Codes of the Sixth Sub-component of LAL	134
Table 4.13.	Codes of the Seventh Sub-component of LAL.....	139

LIST OF FIGURES

Figure 3.1. The development of each semi-structured interview.....	47
Figure 3.2. The processes of making document analysis.....	51
Figure 3.3. The summary of the data collection procedure	57
Figure 3.4. The process of content analysis.....	58



ABBREVIATIONS

ACA	: American Counselling Association
ACL	: American culture and literature
AfL	: Assessment for learning
AFT	: The American Federation of Teachers
ALI	: Assessment Literacy Inventory
EAP	: English for academic purposes
EFL	: English as a foreign language
EL	: English linguistics
ELL	: English language and literature
ELT	: English language teaching
ES	: Educational sciences
ESL	: English as a second language
ESP	: English for specific purposes
ETI	: English translation and interpretation
I1	: Interview 1
ITC	: International Test Commission
I2	: Interview 2
JCTP	: Joint Committee on Testing Practices
LAL	: Language assessment literacy
LTA	: Language testing and assessment
LTAQ	: Language Testing and Assessment Questionnaire
MAC	: Michigan Assessment Consortium
MOE	: The Ministry of National Education
NCME	: The National Council on Measurement in Education
NEA	: The National Education Association
TAP	: Think-aloud Protocols

CHAPTER ONE

1. INTRODUCTION

1.1. Introduction

This chapter starts with the background of the study and goes on with the research problem, purpose and research questions. It gives details on the significance of the study and key terms as well as the overview of the dissertation.

1.2. Background of the Study

Language assessment and testing have been influenced by four approaches. These approaches evolved from several principles. These are as follow: essay translation, psychometric-structuralist, pragmatic and communicative approaches in a chronological order (Alduais, 2012). These approaches evolved from (a) being pre-scientific to being scientific owing to having certain procedures and principles to follow in test design and preparation, (b) intuition to psychometrics and to communication in test preparation and design, (c) assessing language forms separately to assessing language skills in an integrated way, (d) focusing on language forms to focusing on language use including context, meaning and language functions and (e) viewing language as systematically acquired habits to viewing it as a social entity (Alduais, 2012).

Psychometric-structuralist and communicative approaches dominate language assessment and evaluation today (Inbar-Lourie, 2008a). In accordance with these two approaches, testing and assessment culture have been formed with their unique features. As Inbar-Lourie mentioned, the psychometric-structuralist approach has an effect on testing culture because language is the accumulation of tiny pieces of information and the social aspect of language assessment is ignored in testing culture. She also added that the psychometric features of language tests including validity and reliability are paid attention in testing culture. Thus, testing culture aims at assessing a language

learner through summative assessment, which measures what is learnt at the end of a course (Boraie, 2012; Witte, 2010). According to Green (2002), selected-response is commonly used in testing culture to rank language learners since testing culture is. Therefore, the washback effects of a language test on a language learner, school and society is ignored in testing culture as Inbar-Lourie (2008a, 200b) explained. However, the communicative approach influences assessment culture. Inbar-Lourie stated that assessment culture supports the fact that language is constructed by the members of a society through interaction and communication. The psychometric features of language assessment is not significant for it, therefore. Besides, assessment culture recommends that different types of assessments methods should be used to support the language learner during his language learning (Inbar-Lourie, 2008a). According to Inbar-Lourie, the language learner can improve his language learning because his teacher can have a chance to him because assessment culture pays attention to the washback effects of language assessment.

As a result of these changes, the language teacher had a more important place in any educational system, which increased the significance of his assessment beliefs, practices and views. Consequently, the need to standardize language assessment and evaluation arouse. Meanwhile, the American Federation of Teachers (AFT), the National Council on Measurement in Education (NCME) and the National Education Association (NEA) (1990) made the first standardization study to improve the assessment and evaluation practices of the American teachers. Seven standards to follow in assessment and evaluation were developed. According to AFT and its partner organizations, the teacher should (a) choose his assessment method according to the instructional purposes, (b) develop his assessment tools according to the decisions on instruction and (c) administer his exams, score them and evaluate the assessment data. In addition, he should (d) use the assessment results in decision-making related to instruction, students, school and curriculum, (e) develop a valid grading procedure, (f) communicate his assessment results with students and administrators and (g) recognize illegal and unethical assessment practices (AFT et al., 1990). Then, the chief purposes of the standards are to demonstrate that assessment and evaluation are the significant components of education and to indicate that having a good education is not possible without good assessment and evaluation. These seven standards were conceptualized as

assessment literacy in the literature and the literacy requires teachers being master a number of skills (see Stiggins, 1995; Mertler & Campbell, 2005).

Considering the importance of assessment literacy in language classroom, a number of studies show that language teachers have a low or moderate level of language assessment literacy and practice their assessment and evaluation without any training or with little training (Hasselgreen, Carlsen, & Helness, 2004; Taylor, 2009; Vogt, Guerin, Sahinkarakas, Pavlou, Tzagari, & Afiri, 2008). Like the international studies, the studies carried out in nationwide indicate that language teachers need extra training in language assessment and evaluation (Atikol, 2008; Hatipoğlu, 2015b; Mede & Atay, 2017). The need for extra training in language assessment and evaluation points out that language teachers have difficulty in implementing language assessment procedures (designing, administering, interpreting, using and reporting assessment) because of not having enough knowledge and skills to implement language assessment procedures and of being unaware of the concepts and principles constituting language assessment procedures (Fulcher, 2012; Inbar-Lourie, 2013).

In conclusion, the dominant approaches in language assessment and evaluation, assessment culture, the AFT and its partner organizations' standards, the requirements of assessment literacy and having difficulty in implementing assessment procedures (because of the lack of enough training) demonstrate the importance of language assessment literacy. Therefore, language teachers should be assessment-literate to assess and evaluate their students appropriately, effectively and efficiently.

1.3. Statement of the Problem

Teachers are accepted as the key elements for an effective and efficient instruction. Therefore, it is important that they should be assessment-literate. However, assessment literacy studies indicate that the teachers from different majors have generally low or moderate levels of assessment literacy. Like the findings of assessment literacy studies, language assessment literacy studies indicate that language teachers have low or moderate levels of literacy in language assessment and evaluation, which shows that language teachers do not have enough training about assessment and evaluation, cannot make assessment effectively and efficiently and cannot use their

assessment data to improve their students' learning, instruction, curriculum and schools. In spite of their low or moderate levels of literacy in language assessment and evaluation, language teachers assess and evaluate their students in their language classes. Yet, the researcher could not meet any study about the implementation of language assessment literacy in the related literature. Besides, Turkish EFL teachers/instructors assess and evaluate their students in their English classes despite their low or moderate levels of language assessment literacy, but how they implement their language assessment literacy in their English classes is not known as this has not been searched enough in detail in Turkey. Besides, English is taught by Turkish EFL teachers/instructors who have graduated from ELT, English language and literature (ELL), American culture and literature (ACL), English linguistics (EL) and English translation and interpretation (ETI) departments. However, the researcher could not meet any study which indicates how Turkish EFL teachers/instructors graduating from different departments improve themselves in language assessment and evaluation and how their self-improvement affects their assessment and evaluation in their English classes in the related literature.

Considering these issues and the gap in the literature, it was aimed to to investigate to what extent Turkish EFL teachers/instructors are language-assessment-literate and how this affects their teaching in their English classes.

1.4. Purpose of the Study

The main goal of this study is to reveal the implementation of language assessment literacy (LAL) by Turkish EFL instructors in their English classes. It also intends to demonstrate the implementation of its sub-components in English classes.

1.5. Research Questions

To reach the main goal of the study, the following questions guided the study.

1. How do Turkish EFL instructors implement the sub-components of LAL in the class?
2. Which factors affect Turkish EFL instructors' implementation of LAL in the class?

3. What are the effects of LAL on Turkish EFL instructors?
4. Do Turkish EFL instructors encounter any difficulty while implementing LAL? If so, what are they? How do they overcome them?
5. Do Turkish EFL instructors implement all sub-components of LAL? If not, which sub-component is it or which sub-components are they? What causes it/them?

1.6. Significance of the Study

The literature review indicates that language assessment literacy has been studied in several aspects including language teachers' need for training, assessment knowledge base, pre-service assessment training and professional development. Yet, the implementation of language assessment literacy has not been studied. In addition, the researcher experienced several problems like not following the standard criteria in grading writing and speaking exams in language testing and assessment when he worked as a English language instructor and Testing Office member. These affected language assessment and evaluation negatively. When he talked with his colleagues working at other universities, he found out that such problems were common.

Considering the quite little literature regarding language assessment literacy, the researcher's personal experiences and the common problems in Turkey, there is an urgent need to analyze the implementation of language assessment literacy by language teachers in their classes. Then, this study is important for language education training because it can give the instructors of language testing and assessment courses in the Turkish universities feedback about how a language assessment and evaluation course can affect EL teachers, so the instructors can improve their courses in terms of theory, content and practice by seeing the possible effects of their instruction on their student teachers' assessment practices which they will use when they start to work. Besides, similar studies to demonstrate the implementation of language assessment literacy by the EL teachers in primary, secondary, and high schools may be carried out by following the research method of this study.

1.7. Definitions of the Key Terms

The key terms of this study are assessment for learning, assessment literacy and language assessment literacy. Their definitions are explained below.

- 1. Assessment for learning:** Assessment for learning (AfL) is assessing a student's learning during teaching with different assessment methods to improve his learning (Boraie, 2012; Stiggins, 2005a, 2005b; Stiggins & Chappuis, 2006).
- 2. Assessment literacy:** Knowing and being aware of the difference between sound and unsound assessment are defined as assessment literacy (Popham, 2004; Stiggins, 1991, 1995).
- 3. Language assessment literacy:** LAL is the ability which a language teacher needs to have for understanding, analyzing and using his students' assessment data to improve their learning (Inbar-Lourie, 2008a).

1.8. Overview of the Dissertation

This dissertation has six chapters and the first chapter is a quick summary of the dissertation. Introduction is the first chapter. It presents the purpose and research problems of the dissertation.

The second chapter is Literature Review. It explains theoretical framework of the study, assessment literacy and language assessment literacy. Then, it mentions language assessment literacy studies made around the world and in Turkey and the possible contribution of this research to the related literature.

The third chapter is Methodology. It details and elaborates the research design, participants, data collection tools, data collection procedure and data analysis.

The fourth chapter is Findings. It presents the findings obtained through cross-case analysis.

The fifth chapter is Discussion. Findings are discussed depending on the related literature.

The sixth chapter is Conclusion. It mentions the main findings of the study. Second, the chapter expresses the study's implications, limitations and strengths. Then it gives some suggestions for further research.

1.9. Conclusion

This chapter provided detailed information about the background of the study, the research problem and the purpose of the study. The next chapter presents the literature review related to assessment literacy, the standards of assessment literacy and language assessment literacy.



CHAPTER TWO

2. LITERATURE REVIEW

2.1. Introduction

This chapter first explains the theoretical framework of the study. Then, it presents the review of the literature related to language assessment literacy by connecting it with assessment literacy. It also presents the standards of assessment literacy and the studies related language assessment literacy around the world and in Turkey.

2.2. Theoretical Framework: Assessment for Learning

Assessment for learning (AfL) is an instructional intervention in classrooms made by using assessment continuously during the instruction, but not at the end of the instruction (Black, Harrison, Marshall, & William, 2003; Boraie, 2012; Stiggins, 2005a, 2005b, 2006; Stiggins & Chappuis, 2005, 2006; Stiggins & Popham, n.d.). Thus, its aim is to make changes in day-to-day classroom assessment as a teaching and learning process in order to improve a student's learning, but not to measure to what extent he has learnt as a result of instruction (Deluca, Luu, Sun, & Klinger, 2012; McDowell, Sambell, Bazin, Penlington, Wakelin, Wickes, & Smailes, 2006; Stiggins, 2007). Consequently, it provides teachers and students with continuous feedback from formal and informal assessments, so they can have and maintain an ongoing picture of students' learning progress (McDowell et al., 2006; Stiggins & Popham, n.d.).

AfL is student-centered because students have more active roles in assessment and participate into assessment process from the beginning of the instruction to the end; consequently, they use their assessment results actively as the data-driven decision makers of their own learning (Lysaght, 2015; McDowell et al., 2006; Stiggins, 2002, 2006, 2007; Stiggins & Chappuis, 2005, 2006). Therefore, AfL requires a partnership between a teacher and his students. This partnership gives different responsibilities to

the teacher and students. The teacher is responsible for sharing his achievement targets with his students, indicating them the satisfactory and unsatisfactory examples of student work and providing them with feedback (Black et al., 2003; McDowell et al., 2006; Stiggins & Chappuis, 2005). His students are supposed to be more autonomous and proficient by examining the satisfactory and unsatisfactory examples of student work, self-assessing their own work and dealing with the feedback given by their teacher (McDowell et al., 2006; Stiggins, 2006). Thus, the students are expected to make decisions about their learning and implement their decisions by using the continuous information which self-assessment and descriptive feedback provide so that they can benefit from their assessment data to improve their learning. (McDowell et al., 2006; Stiggins, 2005a; Stiggins & Chappuis, 2005; Tulgar, 2017).

Besides, AfL supports students emotionally in a positive way because it indicates that students can achieve success and become successful if they go on studying (Deluca et al., 2012; Lysaght, 2015; Stiggins, 2007; Stiggins & Chappuis, 2005). Emotional support is closely related to academic efficacy and eagerness to learn which are the two focuses of AfL (Stiggins & Popham, n.d.). Academic efficacy is the perceived ability of a student to become successful and have a control over his academic achievement, which makes him ready and willing to learn because the student knows how he can be successful in a given assignment, believes that he can reach success and experiences it in class (McDowell et al., 2006). That is, academic efficacy and eagerness to learn are positively correlated with each other.

Academic efficacy and eagerness to learn demonstrate that each student can learn if they are supported appropriately by the teacher (Stiggins & Popham, n.d.). Therefore, on-going assessment in class is a necessity for AfL because on-going assessment enables the teacher to have a full picture of his students' learning during his teaching. As a result, he can take necessary precautions and intervene with his teaching to improve their learning in cooperation with the students (Black et al., 2003; McDowell et al., 2006; Stiggins, 2002, 2006; Stiggins & Chappuis, 2005, 2006; Stiggins & Popham, n.d.).

In addition, AfL enables the teacher to find out the weaknesses in his instruction and to take precautions to deal with them, so he can improve his students' learning and

his instruction (Lian, Yew, & Meng, 2014, McDowell et al., 2006). The teacher does not grade his assessments, but uses them as a trigger to improve his students' learning and determine how effective his instruction is, so his assessments enable him to rethink and reshape his teaching method, content and activities (Lian et al., 2014; McDowell et al., 2006).

The features of AfL do not want the teacher to use his assessment for giving grades to his students (Lian et al., 2014). Instead, it warrants the teacher to use his assessments for purposes such as gathering information about his students' learning, giving them information about where they are now in their education and helping them to do better the next time (Black et al., 2003; McDowell et al., 2006; Stiggins, 2006, 2007; Stiggins & Chappuis, 2005, 2006).

2.3. Assessment Literacy

Assessment literacy (AL) is that a teacher knows how to assess what his students have learned and how to interpret and use his assessment data to enhance his students' learning and his instruction (Webb, 2002). The definition shows that AL is closely related to the theoretical and practical aspects of assessment and evaluation. According to Popham (2011), the teacher should understand what the reliability of an assessment tool is, know assessment concepts and procedures like reliability and validity and be aware of the concepts and procedures that influence making educational decisions in the theoretical aspects of assessment. In the practical aspects, the teacher should design his assessment in collaboration with his colleagues by choosing and developing assessment methods which measure directly what students are going to learn (Braney, 2011). Then, the teacher should use his assessments by administering them to have a comparable measure of his students' learning, so he can interpret his assessment data to understand the data and feedback his students give, make plans for his instruction and evaluate his instruction (Braney, 2011).

Being assessment-literate requires that the teacher should be critical in dealing with the theoretical and practical aspects of assessment and evaluation effectively and efficiently. Therefore, Abell and Siegel (2011) claimed that it is essential to be aware of the relationship between the teacher's view of learning and assessment knowledge.

According to them, his view of learning is based on his views of how students learn better and what works well in assessment, so the views form the basis of his assessment values and principles which shape and determine his assessment knowledge including the purposes of assessment, assessment interpretation and action taking.

In short, AL is what the teacher should have to use in his classroom assessments: assessment-related knowledge, skills and competencies (Kahl, Hofman, & Bryant, 2013). According to North Central Regional Educational Laboratory (n.d.), it indicates to what extent the teacher is ready to design and implement his assessment and discuss its results. In addition, Mercurio (2013) stated that AL can show to what extent the teacher can understand and reflect on his assessment practices in different practical and theoretical contexts. Thus, he needs to have enough knowledge about the key concepts of assessment and evaluation and to understand the influences of the key concepts on the design of an assessment and the decisions made depending on it (Rogier, 2014).

2.3.1. Why is assessment literacy important?

Teachers should obtain sound classroom assessment data to make sound decisions about their students' learning to provide the students with sound instruction and to benefit from sound decisions (Stiggins, 1991). Therefore, they should distinguish sound data from unsound data. To do so, they should have clear purposes and achievement targets, choose proper assessment method(s), sample their students' achievement and avoid bias and distortion (Stiggins, 1995). They can achieve these by answering why, what and how they will assess, how they will announce and communicate their assessment results and how they will involve their students in assessment (Stiggins, 2006).

Understanding the importance of sound assessment reflects the change in the function of schools that has shifted from ranking students to helping every student to succeed, so this shift has made having the clear meaning of academic success and doing an effective assessment of student academic success essential for schools (Stiggins, 1995). As a result, schools have understood the importance of classroom assessments and of their teachers as teachers are the ones who perform classroom assessment to make their education high quality and sound (Stiggins, 1991, 1995). To achieve this, the

new types of assessment methods have started to be used in schools, but according to Stiggins (1995), teachers have problems in using them because of their misunderstanding, assessment preferences and not being ready to use those.

In addition, the shift in the function of schools has made accountability more important in teachers' lives because authorities and societies determine whether teachers perform satisfactorily depending on their students' assessment results (Mercurio, 2013; Popham, 2011). Therefore, Popham (2011) said teachers should have an understanding of educational assessment because this enables them to have dependable data about their students' learning and to make their instruction more effective (Stiggins, 2014).

In addition to the shift in the function of the school and accountability, teachers' pre-service and in-service assessment training affects their assessment and evaluation practices. Some studies show that pre-service assessment training improves pre-service teachers' assessment literacy (i.e. DeLuca, Chavez, Bellara, & Cao, 2013; DeLuca & Klinger, 2010; Karaman & Şahin, 2014; Lomax, 1996; McGee & Colby, 2014; Richardson, McGee, & Colby, 2015). However, it may fail to prepare pre-service teachers to assess and evaluate their students in their classes (Hofman, & Bryant, 2013; Kahl et al., 2013; Koh & Velayutham, n.d.; Mertler, 2003; Popham, 2006, 2009; Sever & İflazoğlu Saban, 2015; Stiggins, 1991, 1995; Webb, 2002). According to Lomax (1996), this situation may result from the course teachers who are not responsive to the needs of pre-service teachers due to their course content and materials. That pre-service teachers can not practice what they learn may also cause this issue (Rogier, 2014). In addition, Stiggins (1991) told that pre-service assessment courses may be too theoretical, have a narrow perspective and may be neglected. Besides, these courses may be very difficult for pre-service teachers to comprehend because they may include high technical and quantitative standards which may not be very applicable in the classroom (Popham, 2004; Stiggins, 1991, 1995). Consequently, pre-service teachers may not be knowledgeable enough about assessment, so they cannot understand the significance of assessment in improving their students' learning as Popham (2011) mentioned. Besides, they may not cope with the complex challenges of classroom-based assessment and also relate their assessment to their day-to-day classroom practices (Kahl et al., 2013; Popham, 2006; Stiggins, 1995).

Due to inefficient pre-service training, in-service teachers may need extra assessment training (Adanalı & Doğanay, 2010; Akdağ & Ekmekçi, 2015; Cansız Aktaş & Baki, 2012). Some studies show that in-service training increases the levels of in-service teachers' assessment literacy (i.e. Alkharusi, Aldhafri, Alnabhani, & Alkabani, 2012; Engelsen & Smith, 2014; Fan, Wang, & Wang, 2011; Koh, 2011; Koh & Velayutham, n.d.; Q'Sullivan & Johnson, 1993; Mertler, 2009; Volante & Melahn, 2015). Despite this result, schools fall behind on organizing such training programs (Cansız Aktaş & Baki, 2015; Stiggins, 1995) or may organize training programs whose content is not sufficient to help in-service teachers assess their students in an effective way (Adanalı & Doğanay, 2010; Kuran & Kanatlı, 2009; Stiggins, 1995). In-service teachers have different responsibilities that are time-consuming for them to do at school, so in-service teachers cannot find time to implement what they learn in their in-service training (Stiggins, 1995). Koh and Velayutham (n.d.) told that in-service teachers cannot evaluate the quality of their own assessment practices as a result. In addition, the lack of assessment knowledge affects their choice and use of assessment methods. For example, they use either an assessment method they are familiar with (Altun & Gelbal, 2014), or an assessment method they misunderstand and are not ready for using (Stiggins, 1995). Lack of knowledge also restricts their use of assessment results with deciding who passes and fails and whether their students have learned (Eğri, 2006).

In addition to pre-service and in-service training, a teacher's knowledge of assessment tools, view of learning, action taking and knowledge of assessment interpretation also affect his assessment actions (Gottheiner & Siegel, 2012), but Siegel and Wissehr (2011) claim that the teacher's view of learning is more effective in this process. In addition, his teaching approach, assessment value and beliefs, teaching context and content knowledge have an effect on his assessment practices (Izci & Siegel, 2014). External factors like time, materials, workload, the number and level of students and curriculum also influence his assessment actions (Aydoğmuş & Çoşkun Keskin, 2012; Ataman & Kabapınar, 2012; Eğri, 2006; Kuran & Kanatlı, 2009; Özer & Karaoğlu, 2017; Sever & İflazoğlu Saban, 2015). Besides, the way the teacher was assessed when he was a student can determine his attitudes toward different assessment methods and may discourage him from improving himself in assessment (Izci & Siegel, 2014; Stiggins, 1995).

The argumentation during this part indicates that the teacher has to deal with several things in order to assess his students. Therefore, it is essential for him to understand importance of classroom assessment because classroom assessment enables him to evaluate the quality of his teaching (White, 2009; Witte, 2010) and improve his students' learning (Braney, 2011; Smith, Worsfold, Davies, Fisher, & McPhail, 2013; White, 2009; Witte, 2010). It helps him to understand what assessment is and is not and what it can and cannot do (Leighton, Gokiart, Cor, & Heffernan, 2010) and to check whether his students achieve his course objectives (Witte, 2010). Therefore, classroom assessment has a central role in teaching. However, the problems like the lack of enough knowledge influence classroom assessment negatively. It causes the teacher to underestimate the function of classroom assessment in his teaching because of the belief that assessment only measures students' learning, so it is not used for improving the students' learning emphasized by White (2009). As a result, the teacher designs an assessment system which is not appropriate to his students, excludes his students from assessment process and which makes understanding the system difficult for his students (White, 2009). Therefore, his students cannot understand and know why they are assessed, what is assessed and is essential to learn and how they can enhance their learning by receiving feedback. Besides, the lack of knowledge causes the teacher not to connect what he assesses with what he teaches well (Koh & Velayutham, n.d.) and not to have valid and reliable tests (Mertler, 2000). As Bracey (2000) and Popham (2006) stated, the teacher might misinterpret and misuse his assessment data. In addition, the lack of knowledge makes the teacher use a new type of assessment with his students with little information (Lomax, 1996), so he feels unprepared or uncomfortable while assessing his students, which affects the way he prepares his exams negatively (Rogier, 2014).

To sum up, assessment is an everyday activity for the teacher (Quitter, 1999) and he spends 50% of his teaching time by doing assessment activities (Stiggins, 1991, 2014). The different aspects of teaching like making and guiding decisions about large-group instruction and forming and developing personal instructional programs are under the effect of assessment (Mertler, 2003). Therefore, it is important for the teacher to be assessment-literate in order to shape his instruction and to maximize his students' learning (White, 2009). Thus, Witte (2010) stated that the teacher can have information

about the effectiveness of his instruction and make changes if necessary to help his students achieve his goals and objectives. White (2009) warned that if the teacher does not do so, his students may be demotivated to learn, encounter bad long-lasting effects and cannot shape their own learning processes.

2.4. Standards of Assessment Literacy

The basic standards of assessment literacy were developed by the American Federation of Teachers and its partner organizations in 1990. The standards support “the view that student assessment is an essential part of teaching and that good teaching cannot exist without good student assessment” (AFT et al., 1990, p. 1). The standards are also important to understand assessment literacy because they explain how efficient and effective assessment and evaluation should be made in a class in detail. More and more organizations have updated these standards since 1990 depending on the changing conditions in education. The organizations have developed and prepared standards for, codes of and responsibilities for effective assessment, which is the basis of assessment literacy. These organizations include NCME, Joint Committee on Testing Practices (JCTP), Michigan Assessment Consortium (MAC), American Counselling Association (ACA), International Test Commission (ITC) and Turkish Ministry of National Education (MONE). The researcher analyzed the documents prepared by these organizations and institutions. The standards emphasized by these organizations are as follow:

1. Choosing appropriate assessment methods for instructional decisions,
2. Developing appropriate assessments for instructional decisions,
3. Administering exams, scoring them and interpreting their results,
4. Using assessment results in making decision,
5. Developing valid grading procedures,
6. Communicating assessment results,
7. Recognizing unethical, illegal and inappropriate assessment methods and uses of assessment information.

Considering these assessment standards, all seven standards are important for gaining assessment literacy and teachers should be familiar with the standards.

2.4.1. Choosing appropriate assessment methods for instructional purposes

The purpose, intended test takers, and content and skills to be tested are essential for a teacher in choosing his assessment method (JCTP, 2002). An assessment can be used for description, accountability, prediction and program evaluation (MAC, 2013). According to MAC, the teacher must choose assessment methods according to the clear learning targets that his students can understand. He needs to know that quality assessment is an important part of effective teaching and learning as MAC told and to have enough knowledge about the different kinds of assessment as well as their strengths and weaknesses (AFT et al., 1990). According to AFT and its partner organizations, the teacher should know and understand the criteria related to how to evaluate and choose assessment methods depending on educational plans, so he can consider different factors in using and evaluating assessment methods available to him. Consequently, the teacher can obtain and evaluate information about the quality of assessment types. According to AFT and its partner organizations, he should choose assessment methods depending on administrative appropriateness, technical adequacy, usefulness and fairness because he knows how valid assessment supports his teaching, how invalid assessment data affects his students' performance negatively and which assessment methods are compatible with which purposes.

In order to choose assessment and test, the teacher should check the appropriateness of any assessment method to the content and skills to be tested as JCTP emphasized. Besides, he ought to use technical knowledge to decide how to assess their students and know how different assessment methods can influence making decisions about educational things as AFT and its partner organizations told. According to JCTP, he should evaluate assessment methods through their samples and the documents about the methods like directions, answer sheets and score reports. He ought to have information about test takers, norming and standardization procedures, fairness, the accuracy of scoring procedures and modifications (MAC, 2013).

Finally, the teacher should understand and know whether an assessment is appropriate for the intended test takers in terms of the factors including age, grade level or cultural background and whether the assessment has an accurate scoring procedure (MAC, 2013). As a result, he can be sure that several factors like age, gender or

nationality do not affect the results provided by the selected assessment methods (MAC, 2013).

2.4.2. Developing appropriate assessments for instructional purposes

A teacher should know where the data used for making decisions about students come from in developing assessments: assessment and evaluation; therefore, he has to plan how to collect facilitating information for decision-making about student (AFT et al., 1990). AFT and its partner organizations also mentioned that classroom assessment is a dynamic process and requires the teacher to use teacher-made assessment materials as a result. Thus, he should have the knowledge about the principles used for determining how to use and develop different kinds of assessment in the class and to choose different kinds of assessment relevant to his instructional goals (AFT et al., 1990). In addition, the teacher needs to prepare information about what his assessments are going to measure, how they are going to be used, for whom they are prepared and what their strengths and limitations are (JCTP, 2002). According to JCTP, he should (a) tell the development of his assessments and the selection of the content and skills, (b) give information about the technical quality of his assessments and their administration and scoring procedures and (c) supply the samples of his assessment questions and materials to be used in assessing his students. He needs to avoid offensive content and language while preparing his assessment materials (JCTP, 2002). MAC (2013) suggested following a five-step process in developing assessments: “plan, develop, review and critique, field test and review and revise” (p.3). According to NCME (1995), the teacher needs to be sure that he develops his assessment products by preventing bias stemming from the factors including gender, nationality and race to meet the technical, professional and legal standards. He should prepare the documents about how he develops, scores and analyzes his assessments in terms of validity and reliability and how his assessments’ results will be reported as NCME told. NCME also told that the teacher needs to take into account the rights of test takers and the copyrighted materials in developing assessments and the balance in assessment because according to MAC different test users may want to use assessment data for different purposes and different purposes may require the use of different assessment tools.

2.4.3. Administering exams, scoring them and interpreting their results

Giving information to test takers about an assessment, its directions, appropriate test-taking strategies and types of questions and informing them about the consequences of taking and not taking an optional test are necessary to apply assessment methods properly (JCTP, 2002). According to JCTP, a teacher should tell if his students can have the copy of his assessment, how they can retake the assessment and ask for checking it and what their responsibilities are during the administration of the assessment. He ought to be aware of the possible effects of administering an exam on reliability (AFT et al., 1990). According to JCTP, the teacher needs to understand and know the established procedures of administrating an exam, so he can follow and obey them in a standardized way. He has to ensure the security of exam materials, take security measurements during the whole administration process (JCTP, 2002; NCME, 1995) and seek to prevent anything which may invalidate his exam scores (MAC, 2013). Besides, NCME emphasized that the teacher had better protect his students' rights, allow them to ask questions about the exams and directions according to the standardized administration procedure and try to avoid any action which may misrepresent their actual levels.

Providing consistent assessment results is important in scoring exams (AFT et al., 1990). Therefore, the teacher should correct the errors that may affect interpreting scores negatively and report the corrections directly (JCTP, 2002). He also needs to be sure of the confidentiality of the scores through the procedures and by preventing unauthorized release and access as JCTP suggested. Besides, the teacher has to give his students information about what they are supposed to do for the issues related to withdrawing scores (JCTP, 2002). He should control the accuracy of the scores when his students challenge their scores by conducting reasonable quality control procedures before, during and after an assessment (MAC, 2013; NCME, 1995). Besides, according to NCME, the teacher needs to seek to lessen the effects of factors irrelevant to the purposes of assessment on scoring, (b) ensure the confidentiality of his assessment results and (c) develop a reasonable and fair procedure for his students to ask for rescoring.

Understanding the theoretical and conceptual basis of assessments and their procedures, their limitations and the use of scales are essential for interpreting

assessment results (ITC, 2001). The teacher needs to interpret his formal and informal assessment results (AFT et al., 1990) by taking into consideration the norms, content, benefits and limitations of assessment results, modification, technical advice and procedures for setting passing score and performance standards (JCTP, 2002). He should reveal the strengths and weaknesses of his students by using the analysis of the results that is the combination of the information coming from different sources since AFT and its partner organizations, JCTP and MAC stated multiple assessment results can provide a more balanced evaluation of a student. In addition, AFT and its partner organizations told that the teacher ought to find out the reasons for any discrepancy to resolve uncertainty before making a decision if the results are inconsistent and use his assessment results to support his students' learning progress and to prevent their anxiety as well. The use of assessment results for other purposes rather than its intended purpose should be avoided (JCTP, 2002) and the interpretation of assessment results should be considered as giving feedback (MAC, 2013). According to MAC, he should also know (a) the psychometric factors related to validity, reliability, norms and measurement error, (b) factors related to his students including their background, age and gender and (c) the contextual factors like the opportunity to learn, work environment and the quality of educational program. Besides, ITC mentioned that the teacher had better consider any variation from the standardized procedure that may affect assessment results.

2.4.4. Using assessment results in making decisions

Accumulating assessment information is essential for a teacher to make instructional decisions at several levels. Then, the teacher can organize and develop a sound instructional plan to facilitate his students' educational development (AFT et al., 1990). Besides, it is suggested that he should interpret assessment results correctly by preventing misunderstanding and know how to use the results of different assessments appropriately for enhancing his students' learning. In addition, the teacher needs to self-assess his exams and guide instruction depending on assessment results (MAC, 2013).

2.4.5. Developing valid grading procedures

A valid grading procedure is developed depending on students' assessments (AFT et al., 1990). According to MAC (2013), it is a professional judgment and is not a numerical and mechanical exercise. AFT and its partner organizations told that a grading system should be developed by devising, using and explaining a procedure for the development of the grades coming from different assessments. In addition, a teacher needs to identify and avoid any faulty grading procedure, to defend why his grade system is fair, rational and justified and to mention that his grade system reflects his preferences and judgments (AFT et al., 1990). The teacher should evaluate and enhance his grading procedures to improve the validity of the interpretations made about his students depending on the procedures (AFT et al., 1990).

2.4.6. Communicating assessment results

Mentioning the intended interpretation and use of assessment results to students and administrators is significant to communicate assessment results effectively (JCTP, 2002). AFT and its partner organizations (1990) proposed that a teacher should also know assessment terminology and explain the limitation, meaning and implication of his assessment results. The teacher should (a) know and explain the limitations of different formal and informal assessments and (b) understand the significance of measurement errors and consider measurement errors before decision-making (AFT et al., 1990). He ought to communicate assessment results to his students and administrators and demonstrate how his students' progresses are assessed in a timely and understandable manner (AFT et al., 1990; JCTP, 2002). According to MAC (2013), this process enables the teacher to provide descriptive, actionable and timely feedback to his students, so they can use their assessment data to enhance their learning. He should pay attention to his students' background to explain appropriately the interpretations of his assessment results and talk about the printed reports at different levels (AFT et al., 1990). MAC emphasized that he should (a) involve his students in using their assessment results to enhance their learning, (b) give background information about his assessment reports and (c) show his students and administrators how they should interpret assessment results to avoid misunderstanding and

misinterpreting the results. Besides, the teacher had better inform his students and administrators about the effects of his assessment results on them and try to avoid the misinterpretations and misuses of his assessment results as NCME (1995) explained. He also should give oral and/or written feedback to his students and administrators in a supportive and constructive manner (ITC, 2001).

2.4.7. Recognizing unethical, illegal and inappropriate assessments and uses of assessment information

A teacher needs to know (a) that fairness is related to all participants of assessment, (b) what his ethical and legal responsibilities are in assessment and (c) how such responsibilities influence his instructional practices (AFT et al., 1990). In addition, the teacher should know the limits of appropriate professional behavior and the misuse and overuse of different types of assessment so that he can avoid using inappropriate assessment methods and having harmful results as AFT and its partner organizations told. According to NCME (1995), he ought to take into account the confidentiality and privacy of his students and know their rights.

2.5. Language Assessment Literacy

Language assessment literacy as the ability a language teacher should have for understanding, analyzing and using his students' assessment information to improve their learning (Inbar-Lourie, 2008a). It also means that the teacher understands, acquires and masters the skills, knowledge and principles of test construction, interpretation, test use, evaluation, impact and classroom-based language assessment with a critical understanding of how language assessment functions in any educational context (Lam, 2015; O'Loughlin, 2013). Therefore, it includes the knowledge, understanding and practices related to language assessment through which the teacher has to understand, create, analyze and evaluate his assessments in the class (Fulcher, 2012; Malone, 2013; Pill & Harding, 2013; Scarino, 2013). In other words, it is the language teachers' ability to "design, develop and critically evaluate tests and other assessment procedures, as well as the ability to monitor, evaluate, grade and score assessments on the basis of theoretical knowledge" (Vogt & Tsigari, 2014, p. 377). It also includes understanding

the social, historical, philosophical and political frameworks explaining how assessment practices have been developed and how assessment may influence individuals, institutions and society (Fulcher, 2012). According to Malone (2008), language assessment literacy is shortly what the language teacher should know about language assessment.

Most of these definitions are developed depending on the framework Brindley (2001) proposed for preparing a professional language assessment programs as the framework is viewed as the basis of language assessment literacy (Inbar-Lourie, 2008a, 2008b). Brindley told that the language teacher should be trained in five areas: (a) the social context of assessment, (b) defining and describing proficiency, (c) constructing and evaluating language tests, (d) assessment in the curriculum and (e) putting assessment into practice.

Brindley's framework made three questions, why, what and how to assess central to understand language assessment literacy (Inbar-Lourie, 2008a). According to Inbar-Lourie, the first question expresses the rationale of assessment; the second question requires familiarity with the modern theories of learning and assessment and language teaching pedagogy to describe and decide the trait to be assessed; the how of assessment indicates how the language teacher develops appropriate assessments for the evaluation of the trait. Inbar-Lourie (2013) explained that the answers to these three questions form the unique knowledge base which is a combination of general educational assessment principles and language teaching knowledge. General educational assessment principles require being knowledgeable in using summative and formative assessment, in establishing reliability and validity and in interpreting students' scores, while language teaching knowledge includes familiarity with language education, ethicality, applied linguistics and fairness (Inbar-Lourie, 2013).

In addition, Scarino (2013) mentioned that language assessment literacy includes learning theories and their practices, the knowledge of language assessment, curriculum, culture and theories of language. It also encompasses institutional contexts and language teachers' beliefs, experiences and knowledge as well as the contextualized knowledge of language teaching and learning because it is developed through the realities of the language teacher's assessment beliefs, values and experiences in addition

to their practice contexts (Scarino, 2013). Therefore, all of them constitute language teachers' language assessment literacy.

2.5.1. Why is language assessment literacy important?

Language assessment literacy is important for language teachers for four main reasons: language teachers as the agents of assessment, language assessment training, assessment and testing cultures and educational and political reforms. This sub-heading explains them in this order.

2.5.1.1. Being the agent of language assessment

Language assessment affects language teachers' instructional practices and their students' learning processes. Thus, it is considered an integrated and significant part of language teaching, so its integration with language teaching is believed to help students enhance their language learning (Malone, 2013; Rea-Dickins, 2004). As a result of its central role, language teachers are considered as the agents of language assessment (Rea-Dickins, 2004).

Being the agent of assessment makes language teachers responsible for every assessment-related activity such as test preparation, development, administration, scoring and interpretation (Alas & Liiv, 2014; Boyd, 2015; Davison & Leung, 2009; Newfields, 2006; Pill & Harding, 2013). The language teachers are also supposed to identify good and bad assessments as well as the positive and negative effects of their assessments (Boyd, 2015).

Rea-Dickins (2004) said that language teachers' assessment activities are significant to language teaching and learning because language teachers can observe their students. This observation includes assessing their students' performances with different assessment methods. The data this observation provides serves as a basis for language teachers to make decisions about their instructional practices and students' learning progress (Davison & Leung, 2009; Herrera & Macias, 2015; Montee, Bach, Donovan, & Thompson, 2013; Rea-Dickins, 2004). Besides, Davison and Leung pointed out that teacher-based language assessment is dialogical as students learn language through the guidance and advice language teachers give to them to enhance

their language learning. Therefore, language teachers can monitor their instruction and adjust it if necessary through language assessment, while students can monitor and improve their learning via language assessment (Herrera & Macias, 2015).

On the other hand, several factors related to them like their knowledge of second language learning, assessment and student learning can prevent language teachers from doing language assessments efficiently and effectively. If the language teachers do not possess a sound assessment knowledge base, they may encounter several problems in their assessment practices. These problems may include: (a) not understanding the importance of classroom-based language assessment in language teaching (Shohamy, Inbar-Lourie, & Poehner, 2008), (b) limiting the use of classroom language assessment to giving grades by ignoring its implications and disintegrating language assessment and teaching (Herrera & Macias, 2015), (c) having misconceptions about the types of assessment methods (Davison & Leung, 2009) like considering one type superior to other types (Lam, 2015) and (d) causing validity and reliability problems by not administering and marking exams appropriately (Alas & Liiv, 2014). These problems may cause the language teachers to make wrong decisions depending on their assessment data and these decisions may affect language teaching and learning negatively (Pill & Harding, 2013). Besides, the language teachers may consider language assessment as a hindrance (Montee et al., 2013) and may use wrong classroom-based language assessment procedures (Rea-Dickins, 2008).

2.5.1.2. Language assessment training

Language assessment training has an important place in language teachers' assessment practices. However, language teachers assess and evaluate their students in their classes with insufficient assessment training or without any assessment training (Hasselgreen et al., 2004; Taylor, 2009; Vogt et al., 2008) because the language assessment training language teachers receive is not comprehensive and their critical awareness in language assessment is not developed enough (Vogt et al., 2008).

Language teachers are trained about language testing and assessment through pre-service and in-service assessment training courses, but these courses cannot be sufficient and efficient. The main reason is that pre-service language assessment

training is not attached enough importance (Lam, 2015; Taylor, 2009). Besides, the non-language experts with an incomprehensive content give pre-service training (Riestenberg, Di Silvio, Donovan, & Malone, 2010; Riazi & Razavipour, 2011). In addition, pre-service language assessment training is considered too theoretical and technical for language teachers to understand (Malone, 2013; Taylor, 2009, 2013) owing to the course materials prepared by professional testing organizations and language education institutions (Davies, 2008; Taylor, 2009). Thus, it does not help pre-service language teachers to improve their assessment and evaluation (Riazi & Razavipour, 2011). Because of insufficient pre-service assessment training, language teachers need in-service assessment training in order to improve their assessment and evaluation practices (Hasselgreen et al., 2004; Vogt et al., 2008). However, language teachers may not benefit from in-service assessment training because of some factors like cost and time (Riestenberg et al., 2010), so they cannot implement what they have learned in their classes (Riazi & Razavipour, 2011).

2.5.1.3. Assessment culture vs. testing culture

Educational theories affect language assessment and evaluation directly by creating two different cultures related to language assessment and evaluation: assessment culture and testing culture. According to Inbar-Lourie (2008b), behaviorism causes testing culture to view language as the accumulation of the small bits of knowledge, so testing culture gives a passive role to students, focuses on the psychometric features of assessment and aims to check what students have learnt in terms of micro-linguistic aspects of language like grammar and vocabulary. On the other hand, assessment culture views language as a social practice, so it gives importance to the communication and interaction between language teachers and students, gives an active role to students in their assessment and aims to improve students' language skills through different types of feedback including self-, peer and teacher feedback (Inbar-Lourie, 2008a, 2008b).

Language assessment and evaluation has shifted its focus from testing culture to assessment culture (Inbar-Lourie, 2008a). This shift requires language teachers to acknowledge the current political and educational ideologies as well as social values, expectations and attitudes (Inbar-Lourie, 2008b). Therefore, they should assess and

facilitate language instruction, mediate their instruction by providing their students with feedback, use multiple types of assessment methods for collecting data about their students' learning progresses and integrate language assessment with language teaching (Inbar-Lourie, 2008a, 2008b).

If language teachers are not familiar with testing and assessment cultures, they cannot handle the context-specific factors including their beliefs about themselves and about the social, institutional and cultural contexts of their assessment practices in their teaching contexts appropriately (Davison, 2004; Riazi & Razavipour, 2011). Davison told that language teachers' beliefs are related to the purpose of language assessment, the relationship between teaching and assessment, their role in language assessment and their previous knowledge about students. In addition, testing and assessment cultures are the reflections of the such factors that directly influence language teachers' assessment practices through their beliefs. Therefore, if language teachers do not know what testing and assessment cultures are, they cannot: (a) understand what their teaching contexts want them to do, (b) take part in making decisions about assessment and evaluation actively, (c) use their assessment results to improve their teaching and students' learning and (d) take initiative in making changes in assessment and evaluation for the sake of themselves and their students (Davison, 2004; Inbar-Lourie, 2008b; Riazi & Razavipour, 2011).

In conclusion, language teachers should be familiar with educational theories. Familiarity with theories can help them to extend their assessment knowledge they are supposed to know and to make necessary changes in their instruction and assessment (Scarino, 2013). Language teachers should know testing and assessment culture, use this knowledge in test development, interpretation and analysis and try the critical perspectives for specific purposes in different contexts while assessing their students (Scarino, 2013).

2.5.2.4. Educational and political reforms

Educational and governmental authorities systematically make educational reforms to enhance students' learning by making changes in educational policies and practices due to lack of teacher competence, poor student performance in international

tests, insufficient learning outcomes and deficient learning standards (Brindley, 2008; Broadfoot, 2005; Duong, Pham, & Thai, n.d.; Inbar-Lourie, 2013; Malone, 2008; Walters, 2010). Educational reforms use testing and assessment as tools to change directly what happens in the classroom and to provide measurable and visible results for accountability (Brindley, 2008; Broadfoot, 2005; Duong et al., n.d.; Inbar-Lourie, 2013; Malone, 2008; Rea-Dickins, 2008). Therefore, assessment and testing are the indispensable parts of educational reforms. Educational reforms also influence language teachers and students directly because reforms determine how language teachers teach and assess and how their students should study to learn the language, yet the effects of this process may be incompatible with the intentions of test constructors and educational reformers (Brindley, 2008).

As educational reforms use testing and assessment to achieve their goals, language testing and assessment influences language teachers to achieve the goals negatively. Language teachers ignore the consequential validity of their assessments on their teaching and students (Broadfoot, 2005). Thus, they reduce the efficiency of their assessment and use one type of assessment more than other types of assessment in order to achieve the goals of educational reforms, which causes one type of assessment to dominate education and limit language teachers and students' capacities and causes language teachers to ignore the affective domain of education (Broadfoot, 2005). Therefore, language teachers do what the authorities making educational reforms want without questioning the possible effects of this like being restricted to a certain type in shaping and constructing their assessments, not developing themselves professionally and limiting the opportunities to evaluate students' language learning development (Leung & Lewkowitz, 2006; Rea-Dickins, 2008; Saad, Sardareh, & Ambarwati, 2013). In addition, language teachers may not implement educational reforms and their standards effectively in the class because they may not be trained about reforms, implement the reforms' requirements in the class and assess some domains of the standards (Walters, 2010). Thus, educational reforms cannot achieve their goals and create long-lasting effects in language teaching and learning (Brindley, 2008).

As understood from the discussion, the center of any educational reform is language assessment (Taylor, 2009). According to Fulcher (2012), language exams are considered as an implication tool of educational systems, so they are used for

controlling language teachers in their classes and holding them accountable for the implementation of educational goals. Therefore, language teachers are considered as the target of the expected and determined effects, yet they cannot resist, change or affect the policy because they have low or moderate levels of language assessment literacy to produce their counter-arguments depending on their understanding and use of language assessment (Fulcher, 2012).

2.6. Language Assessment Literacy Research

Considering the literature, language assessment literacy has gained importance in recent years and an expanding literature is dealing with this issue nowadays. Therefore, the literature related to the purpose of the dissertation have been presented under two sub-headings: the studies made around the world and in Turkey.

2.6.3.1. Language assessment literacy research in the international sphere

Language assessment literacy studies made around the world were investigated under five different, but interrelated categories. These categories include language assessment courses, professional development, language teachers' need for training, their assessment beliefs and practices and their assessment knowledge base. The findings of the studies in these categories were discussed by relating them to each other to indicate how the present study would contribute to the field of language assessment and evaluation around the world.

Different researchers dealt with the different aspects of pre-service language assessment courses in their studies. Two of these studies (i.e., Brown & Bailey, 1996; Bailey & Brown, 2008) showed the basic characteristics of language assessment courses and the change between these courses in 1996 and 2008. These studies indicated that the assessment courses balanced theory and practice and were taught by the instructors more experienced in language testing and assessment in 2008, while the courses were evaluated positively as being interesting and useful or negatively as being too theoretical and difficult by pre-service language teachers in 1996 and 2008. Similarly, Jin (2010) and Jeong (2013) searched the characteristics of language assessment courses in terms of the effects of the course instructors on the courses' content. Jin found out

that the course instructors taught reliability, validity, item writing, item facility and discrimination, score interpretation and testing four skills by integrating theory with practice, but they did not spend enough time practicing the theory. In addition, Jeong demonstrated that the course instructors with language assessment background focused on the theoretical aspects of language assessment while the ones without such background dealt with the practical aspects of language assessment in their courses, which affected the way the instructors chose their course books used in their courses. Like the course instructors, what the language assessment course books focused on in their content varied from skills (appropriate and necessary methodology for things like test analysis, item writing and statistics) to skills and knowledge (background information about measurement and language descriptions) and to skills, knowledge and principles (impact, the proper use of tests, ethics, fairness and professionalism) (Davies, 2008). In addition to these studies on language assessment course instructors and course books, Lam (2015) revealed that teacher education institutes might focus on the theoretical aspects of language assessment more than its social dimensions like validity, impact and fairness, which made their pre-service language teachers incompetent in language assessment in terms of skills, knowledge and principles of language assessment. As a result, a huge disjuncture between language assessment courses at universities and assessment practices at schools occurred (Lam, 2015).

Apart from these studies, some researchers investigated the training that language teachers got in their studies. The studies (e.g., Hasselgreen et al., 2004; Vogt et al., 2008; Guerin, 2010; Vogt & Tsagari, 2014) revealed that the language teachers in the European countries had little pre-service training in three or four areas of language testing and assessment (classroom-focused testing and assessment, content and concepts, purposes of testing, and/or external tests and exams); therefore, they needed extra training. The lack of sufficient pre-service training caused those language teachers to form their assessment knowledge on the job, to implement assessment tools inappropriately in their classes, to have negative experiences and not to evaluate their assessment practices critically (Tsagari & Vogt, 2017; Vogt & Tsagari, 2014). Owing to deficient pre-service training, those teachers could not identify the areas where they wanted to be trained more and assist their students to improve their learning (Tsagari & Vogt, 2017). According to Tsagari and Vogt, the pre-service language testing and

assessment training was neglected and those teachers were dependent on the traditional forms of assessment though they were supposed to apply the non-traditional forms of assessment in their classes. As a result, Fulcher (2012) revealed that language teachers wanted more training in the basic concepts of language testing and assessment like validity, reliability, classroom-based and large-scale testing and washback which integrated theory with practice. In addition, language teachers wanted language assessment course books to include the real-life assessment activities which they might encounter in their classes (Fulcher, 2012).

Insufficient pre-service language assessment training made professional development more important for language teachers. Therefore, some studies (e.g., Mahapatra, 2016; Montee et al., 2013; Nier, Donovan, & Malone, 2013; Riestenberg et al., 2010; Walters, 2010) indicated the importance of online or face-to-face professional development programs to improve in-service language teachers' language assessment literacy and investigated the effects of such programs on the participants' language assessment literacy. Walters (2010) indicated that the participant language teachers started to evaluate their assessment critically by aligning their assessment with the standards set by the governmental organizations so that the participants could meet the expectations of the governmental organizations from them. Nier and her colleagues (2013) also discovered that an online assessment course enabled their participant language teachers to feel more comfortable with many assessment terms by leading to a positive change in the participants' understanding of assessment and their future plan. In addition, according to Riestenberg and her colleagues (2010), an online professional development course enabled their participant language teachers to learn the basics of assessment (purposes of assessment, validity, reliability, practicality and impact) and to apply what they learned to their courses. Besides, Montee and her colleagues (2013) enabled their participant language teachers to be more confident in their assessment practices, to link their assessment with their teaching and to engage their students in their assessment practices more through a short-term face-to-face professional development course. Consequently, professional development improved the participant language teachers' language assessment literacy by helping them to be familiar with assessment terms, to evaluate their assessment more critically, to apply what they learned to their assessment practices, to engage their students into their assessment

practices and to link their assessment with their teaching. Besides, Mahapatra (2016) prepared an online assessment course for his participant language teachers to improve their language assessment literacy. His study showed that the participants could improve their language assessment literacy with the help of web 2.0 tools in the online program. In addition to these studies, Malone demonstrated what the participant language testers and teachers thought about the content of an online language assessment literacy program in her study through which she developed the program. Her study revealed that the participant language testers were more interested in the detailed presentation of the theoretical aspects of language assessment, while the participant language teachers found the clear and concise presentation of the practical aspects more important in the program.

Apart from the studies above, there are some other studies which investigated the language assessment literacy levels of the EFL teachers and teachers teaching English as a second language (ESL) in terms of assessment beliefs. According to Rogers, Cheng and Hu (2007), the participant EFL/ESL teachers from Canada, China and Hong Kong believed that language assessment helped them to improve their instruction and their students' learning because assessment results enabled the participants to focus on their instruction more and assisted their students to learn language better by motivating and giving them different learning opportunities. Yet, there was a disjuncture between the participants' assessment practices and beliefs because they used paper-and-pencil tests though they believed non-traditional assessment methods should be used in language assessment (Rogers et al., 2007). In another study, Shohamy and her colleagues (2008) revealed that the participant language teachers supported teaching pragmatics, metaphor and culture and using alternative and diagnostic assessment in advanced language classes because the participants believed their students were self-motivated to learn, responsible for their learning and self-aware of the importance of learning. Yet, the participants, like the ones in Rogers and his colleagues' research, used summative assessment instead of formative and diagnostic assessment (Shohamy et al., 2008). In a different study by Munoz, Palacio and Escobar (2012), the participant English teachers in an institution believed that assessment could enhance teaching and learning and help to evaluate the performance of an institution; therefore, assessment could affect teaching and should be

used for formative purposes. Despite their beliefs, the participants did not benefit from their assessment results and use their assessment for formative purposes (Munoz et al., 2012). These results were in line with Rogers et al., 2007; Shohamy et al., 2008). Similarly, the main reasons for the disjuncture between assessment beliefs and practices were the number of students, lack of time, standardized tests, great labor (Rogers et al., 2007) and teaching context, experience and lack of training (Shohamy et al., 2008). On the other hand, the participant elementary EFL teachers in Chan's research (2008) followed their assessment beliefs in their assessment practices. The participants in the study believed that assessment was a part of their responsibility, the alternative assessment was more effective and multiple assessments improved their teaching and their students' learning. Unlike the participants in the previous studies, the participant teachers in the study used alternative and multiple assessments in their teaching in order to understand their students' learning achievement and progress and to evaluate the effectiveness of their instruction. Chan also added that work overload and time-consuming activities affected his participants' assessment practices. In addition, Jannati (2015) found out that the participant Iranian EFL teachers shared the same assessment beliefs with the other participants in the previous studies. Apart from assessment beliefs, Jannati's study also dealt with the participants' knowledge about the fundamentals of language assessment like validity, reliability, fairness and washback. The participants knew the fundamentals of language assessment, but they were not familiar with the ways to make their exams valid, fair and reliable, which caused them not to pay attention to those fundamentals in their assessment practices (Jannati, 2015). Jannati added that the course objectives, curriculum, students' language proficiency and their ages influenced the participants' assessment activities. Besides the previous studies on assessment beliefs and perceptions, Hidri (2015) investigated the relationship between the EFL teachers' language assessment literacy and their conceptions of assessment in Tunus. His research demonstrated that improvement, accountability and irrelevance influenced the participants' language assessment literacy. In addition, he revealed that his participants found assessment irrelevant as they were blamed for their students' failure because of not preparing their students for their work life. Hakim also (2015) investigated the EFL teachers' language assessment literacy in a different perspective (ideology). She found out that the participants' teaching experience

determined to what extent they reflected their understanding of assessment concepts in their assessment practices (the most experienced participants reflected their understanding more than the low and moderate experienced ones).

Some other studies focused on language teachers' assessment knowledge and its effects on their assessment and evaluation practices. Kiomrs, Abdolmehdi and Naser (2011) found out that their participant Iranian EFL teachers had low level of language assessment literacy; therefore, their assessment practices were severely affected by the standardized tests because they only knew the standardized tests and believed that such tests were the perfect tools to assess their students, so their exams copied the structure of the standardized tests with or without making small changes. Thus, according to Kiomrs and his colleagues, the participants could not compensate the negative washback effects of the standardized tests. In a similar study, Leaph, Channy and Chan (2015) revealed that the Cambodian ELT instructors used the standardized tests inappropriately in their assessment practices because their language assessment literacy levels were low. According to Leaph and his colleagues, the participant instructors had a low level of language assessment literacy because most of them were not trained about the standardized tests, some of them did not take such tests before and they did not know the difference between the purpose of classroom-based assessment and the purpose of the standardized tests. Similarly, Talib, Kamsah, Ghafar, Zakaria and Naim (2013) found out that the Malaysian language teachers could not meet the requirements of the education reform in the country which required them to be familiar with the basic concepts of language assessment because they had low levels of language assessment literacy. Like the findings of these studies, Xu and Brown (2017) pointed out that the Chinese EFL teachers at Chinese universities had a low level of language assessment literacy because of the lack of assessment policies and professional standards, inadequate pre-service and in-service training and the absence of assessment literacy in recruitment criteria.

In addition to these studies, other studies dealt with the meaning of language assessment literacy for language teachers and the effect of peer work on language teachers' language assessment literacy. According to Razavipour (2014), language assessment literacy was having necessary skills to assess and evaluate students' language development for the participant language teachers. Razavipour also revealed

that the participants generally depended on their own experiences as students to build their language assessment literacy. In order to improve language teachers' language assessment literacy, Tahmasbi (2014) used scaffolding and artifacts as peer work activities to improve the participant in-service EFL teachers' language assessment literacy. According to Tahmasbi, the peer interaction between the participants in the experimental group helped them to improve their assessment literacy by peer assessing their peers' products and applying what they learned to their own products, which made the participants capable of using peer interaction to improve their own and peers' assessment abilities.

The last group of studies indicates how a good level of language assessment literacy affects language teachers' instruction. Having a high level of language assessment literacy enabled language teachers to benefit from assessment-based dialogues in the classroom by knowing that each assessment provided different learning opportunities to students (Rea-Dickins, 2006). According to Rea-Dickins, this, therefore, helped the participant language teachers to provide orientation toward achieving goals and assisted their students to increase their language awareness and understand language knowledge better. In a similar study, Hamp-Lyons (2017) found out in her small-scale and exploratory study that having a good level of language assessment literacy could help language teachers to reveal and turn learning-oriented assessment opportunities into formal tests. In another study, according to Scarino (2017), being language-assessment-literate could enable language teachers to cope with what intercultural-orientated language teaching brought. An intercultural orientation in language teaching created a challenge for language teachers because it required them to re-conceptualize the construct(s) that they would assess and to alter the processes of eliciting evidence of their students' learning and the frames of reference used as context for making judgments about their students' learning (Scarino, 2017). Besides these, language teachers could meet the expectations of national education reforms owing to having a good level of language assessment literacy (Sellan, 2017). Sellan showed that the participant Singaporean language teachers took responsibility and expanded their assessment constructs by caring culture more, extending understanding of genres, paying attention to content knowledge and practicing high-order thinking,

communication and learning in real-life contexts, so they developed their language assessment literacy and improved their students' learning.

To conclude, the studies above have first shown that pre-service language assessment courses have experienced some changes like from being more theoretical to being balanced between theory and practice, the language assessment course instructors have a big effect on the courses' designs and language assessment course books have been prepared under the effects of three trends (skills, skills + knowledge and skills + knowledge + principles). Besides, they have revealed that language teachers do not have enough pre-service training about language assessment and evaluation, so they need more in-service training about this, assess their students without enough training and improve their assessment skills on the job. According to the studies, insufficient pre-service assessment training and need for more in-service training help to develop some face-to-face and online professional development programs which improve language teachers' language assessment literacy. However, some studies have revealed that there is a disjuncture between language teachers' assessment beliefs and practices because of some factors like the number of the students and workload. This part has also indicated how the low and high level of language assessment literacy affects language teachers. However, these studies have not explained how language teachers implement their language assessment literacy in their classes in terms of assessment standards stated by AFT and its partner organizations (1990). In addition, they have not given enough information about how some factors like the number of the students, experience and workload influence their implementation of language assessment literacy in their classes. Therefore, the present study would contribute to the existing literature by giving elaborated and detailed information about the implementation of language assessment literacy by the EFL teachers in their English classes.

2.6.3.2. Language assessment literacy research in the national sphere

The studies related to language assessment literacy in Turkey were investigated under four different, but interrelated categories. These categories include pre-service assessment training, assessment knowledge base, exams prepared and personal factors (beliefs, attitudes and practices). The findings of the studies in these categories were

discussed by relating each section to one another in order to indicate how the present study would contribute to the field of language assessment and evaluation in Turkey.

Research showed that the pre-service language assessment training is incomprehensive and inefficient in the Turkish context. On this issue, Hatipoğlu (2010) indicated in her small-scale descriptive study that the pre-service ELT students had only one language assessment course during under-graduate years which they considered insufficient to learn and practice the issues and concepts of language assessment. The pre-service teachers also expressed that this course did not balance theory and practice and then they could not evaluate their assessment practices critically. Supporting this finding, Hatipoğlu and Erçetin (2016) added that a lecturer cannot cover all of the issues and concepts of language assessment in one course, but just mentions the issues and concepts superficially by hoping that he can increase his pre-service ELT teachers' awareness in language assessment without giving them enough opportunities to practice what they learn in their systematic literature review. In addition, Hatipoğlu (2015a) found in her need analysis survey that the pre-service ELT teachers reflected the effects of the local assessment cultures and contexts and their previous assessment experiences on their needs and expectations from language assessment course, which affected them and pre-service assessment course negatively. Unlike these studies, Yetkin (2015) claimed in his small-scale survey-based research that the pre-service Turkish ELT teachers improved their knowledge of assessment in the pre-service assessment course with the help of their assignments and school practicum course.

As a reflection of incomprehensive and inefficient pre-service assessment course, the language assessment literacy levels of the in-service Turkish ELT teachers working at the different stages of education were found to be low depending on how much they knew about language assessment and how well they were trained (Büyükkarcı, 2016; Hatipoğlu, 2015b; Mede & Atay, 2017; Öz & Atay, 2017; Şahin, 2015). The participants in these studies worked at the state and private primary, secondary, high schools and preparatory departments of the state and foundation universities in Turkey. Their knowledge of assessment was measured by these researchers with two different tools: "Assessment Literacy Inventory (ALI)" (Mertler & Campbell, 2005) and "Language Testing and Assessment Questionnaire (LTAQ)" (Vogt & Tsagari, 2014). Büyükkarcı (2016) used ALI to determine the assessment

literacy of his participants based on the number of the correct answers his participants gave to the questions related to seven different assessment situations in ALI. On the other hand, the other researchers (e.g. Hatipoğlu, 2015b; Mede & Atay, 2017) used LTAQ to determine their participants' level of assessment literacy according to their received assessment training and perceived needs for assessment training in three components: (a) classroom-focused language testing and assessment (LTA), (b) purposes of testing and (c) content and concepts of LTA (Vogt & Tsagari, 2014). No matter how the data were collected by these Turkish researchers, they have found out that the in-service Turkish ELT teachers have limited LTA expertise because of their insufficient knowledge of assessment. In addition, Öz and Atay (2017) indicated that the in-service Turkish EFL teachers were familiar with basic classroom assessment, but there was a difference between their assessment perceptions and practices.

Lack of sufficient assessment knowledge because of deficient pre-service assessment training affects the in-service Turkish EFL teachers' exam preparation negatively. Köksal (2004) and Sariçoban (2011) examined the exams prepared by the in-service Turkish EFL teachers working at the state schools in their document analysis studies. While Köksal found serious problems in the teachers' exams related to timing, scoring, naming sections, spelling, punctuation, readability, the level of students, construct validity, contextualization, instruction, content validity, washback and reliability in 2004, Sariçoban (2011) revealed that the in-service Turkish EFL teachers improved their exams in terms of face validity, spelling, punctuation, instruction, timing, contextualization, scoring, readability and reliability. Yet, the participants in these studies had still problems with content and construct validity, naming sections and washback. Apart from these studies, Kırkgöz and Ağçam (2012) investigated the effect of the curriculum change in the question types that the in-service Turkish EFL teachers used in their exams in the primary schools. They found out that the curriculum change increased the use of constructed response items in the exams in comparison with selected response items.

Though the limited knowledge of assessment affects the attitude of the in-service Turkish EFL teachers toward different types of assessment because of the insufficient pre-service and in-service assessment training (Aksu Ataç, 2012), several researchers (e.g., Büyükkarcı, 2014; Han & Kaya, 2014; Öz, 2014) indicated that pre-

service and in-service assessment training courses do not affect the in-service Turkish EFL teachers' assessment beliefs and practices. For example, Büyükkarcı (2014) showed in his small-scale mixed methods study that even though the in-service Turkish EFL teachers at primary schools had positive beliefs about formative assessment, they did not use it effectively because of the number of the students in their classes and their workload. In addition, assessment of learning or summative assessment affected the in-service EFL teachers' purpose of using assessment and choice of question types. For instance, Öz (2014) showed in his survey-based study aiming to find out the in-service Turkish EFL teachers' practices of assessment for learning that the participants did not want their students to be involved in assessment procedure and did not help their students to improve their weaknesses. Besides, Han and Kaya (2014) demonstrated the effect of the in-service EFL teachers' beliefs about different language skills on how often these language skills were assessed. For instance, their study showed that the participants did not consider listening important, so listening became the least frequently assessed skill. In addition, Gönen and Akbarov (2015) demonstrated in their exploratory study on a classroom-based assessment that the in-service Turkish EFL instructors could not put some of their assessment beliefs into practice because of the centralized assessment system, their syllabi and their students' educational background.

To sum up, the literature reveals that the in-service Turkish EFL teachers assess and evaluate their EFL students in their English classes without sufficient assessment training in different stages of education. In addition to the lack of sufficient assessment training, their assessment practices are influenced by the number of the students, their workload, need for extra training, their beliefs about four language skills, their syllabi, their students' educational background and the standardized assessment system. Yet, the studies do not explain how the in-service Turkish EFL teachers/instructors assess and evaluate their students in their English classes by dealing with the seven stages of language assessment and evaluation stated in the literature. The studies fall behind showing how the in-service Turkish EFL teachers/instructors are affected by some factors like the number of the students and workload in the seven stages of language assessment and evaluation. Therefore, the present study would be valuable as it specifically focuses on the gaps in the literature and presents the implementation of

language assessment literacy by in-service Turkish EFL teachers/instructors in their English classes.

2.6. Conclusion

This chapter has given information about the theoretical framework of the study, assessment literacy, standards of assessment literacy and language assessment literacy. It also deals with the literature related to language assessment literacy research in international and national sphere. The next chapter is going to detail and elaborate the methodology of this study.



CHAPTER THREE

3. METHODOLOGY

3.1. Introduction

This part first explains the research method and design. Second, it gives information about how each participant selected. Third, it mentions each data collection tool and the relationships between them. Fourth, it indicates how this study was made trustworthy. Finally, it ends up with data transcription, collection procedure and analysis.

3.2. Research Design

In this study, a case study research design from qualitative research methodology was used in order to investigate the implimentation of language assessment literacy because the scope of the case study is based on a contemporary phenomenon which is searched deeply in its real context, but the boundaries between the phenomenon and context are not clear (Yin, 2009). In order to design the case study and establish the logic of the case study, Yin stated that the five components of a research design should be known. They are (a) the research questions, (b) the research propositions (if any), (c) the research's unit(s) of analysis, (d) the logic to link the data to the propositions and (e) the criteria to be used in data interpretation (Yin, 2009).

How and why questions are mainly used in the case study (Yin, 2009). In order to determine the research questions of the study, literature was reviewed. The researcher first decided what to study in his dissertation. Then he narrowed down his interest to a key topic through the literature review and by consulting to a language assessment expert who had a Ph.D. degree in ELT and was specialized in language testing and assessment. In order to form his research questions, the research questions of the studies which focused on assessment and language assessment literacies were analyzed.

It is important for a researcher to study within the scope of the study (Stake, 1995; Yin, 2009). Yin believed that a proposition in a study directs the researcher's attention during the study, so he can stay in the feasible units in data collection. In addition, Stake thought having specified research questions enables the researcher to achieve this. The main proposition and sub-propositions of this research were formed based on the standards developed for assessment literacy by AFT and its partner organizations (1990) because language assessment literacy is a term derived from assessment literacy, so it is closely related to assessment literacy. Besides, there are still discussions about the basic components of language assessment literacy (Fulcher, 2012); therefore, the standards by AFT and its partner organizations were used as the propositions which directed the researcher what to include and study in this study. The main proposition is that an EFL instructor should be language-assessment-literate in order to implement his classroom-based language assessment effectively and efficiently. This shows the main focus of the research.

The unit of analysis which may be one thing or a group of things is related to the definition of the case (Patton, 2002). A group of Turkish EFL instructors was used as the unit of analysis because the research questions were asked to find out how language assessment literacy was implemented by different EFL instructors in their English classes. Besides, the questions were asked to understand and reveal the different or similar perspectives, implementation, interpretations, opinions, feelings and attitudes of different EFL instructors in terms of language assessment literacy. Though the unit of analysis is a group of Turkish EFL instructors, each participant of this group is also a sub-unit of the analysis.

The fourth and fifth components contribute to data analysis steps in a case study. If they are determined carefully, they can provide a solid foundation for data analysis (Yin, 2009). The first step of these components is to determine an analytic strategy because it shows which evidence is going to be used to answer the research questions, helps to treat data more fairly, to produce compelling analytic conclusions, to rule out alternative explanations and to use data collection tools more efficiently (Yin, 2009).

The researcher in this study used the first analytic strategy proposed by Yin: relying on the theoretical propositions. According to Yin, this strategy directs the

research questions, objectives, design of the study, data collection plan and literature review, which helps to focus on the certain data and ignore the others, therefore. The literature review of this study shows that language assessment literacy is a new term and that there are still several discussions about its definitions, knowledge base and standards. On the other hand, most of the assessment literacy studies in the literature are based on the standards developed by AFT and its partner organizations. Therefore, the same standards of the AFT and its partner organizations were used as the theoretical propositions to study language assessment literacy in this research. The research questions were also based on them and the data collection tools were determined accordingly.

The second step in the fourth and fifth components is to choose an analytic technique (Yin, 2009). The researcher chose cross-case synthesis or analysis which is used in multiple-case studies, requires the separate analysis of each case and compares each analysis with each other in case study report (Creswell, 2007; Yin, 2009). Each case was analyzed separately and the case study report was written by comparing and contrasting the results with each other.

A case study is designed based on their functions, characteristics and disciplines (Hancock & Algozzine, 2006). In addition to these design categories, Creswell mentioned that it is also designed based on the number of the case. The last design criterion is effective in the categorizations of Stake and Yin. The case study can be instrumental which aims to understand an issue, while a collective case consists of several instrumental cases to understand an issue by combining information from smaller cases (Hancock & Algozzine, 2006; Stake, 1995). Similarly, a single case study is composed of one case, while multiple-case study consists of more than one case like Stake's collective case study (Yin, 2009). Yin also added that there are two types of multiple-case study: a holistic multiple-case study which analyzes the whole units without any sub-unit and embedded multiple-case study that analyzes both the whole units and sub-units together.

This case study was designed as an embedded multiple-case study (collective case study) because it focused on the eight Turkish EFL instructors and how they implemented language assessment literacy in their classes. Each participant in this study

was accepted as a single case and the same research procedure was carried out with each of them separately. That is, each participant was interviewed three times and observed twice individually. Besides, the participants provided documents for document analysis and joined a focus-group discussion. The data collected from each participant first was analyzed separately. Then, they were cross-analyzed together.

To sum up, the researcher in this study followed the ways that Yin recommended using in order to design a case study and establish its logic. How he designed the present research was indicated briefly in Table 3.1 below.

Table 3.1.

The Summary of Designing the Present Study in Five Stages

Yin's Suggestions	What the Researcher Did
1. Using research questions starting with how and why	- Making a comprehensive literature review - Consulting to an ELT language testing and assessment expert - Analyzing the research questions of the previous studies
2. Using research propositions	- Adopting the standards of assessment literacy as the propositions of the study
3. Determining the unit(s) of analysis	- Using 8 Turkish EFL instructors as the unit of analysis owing to the aim of the study - Using each participant as the sub-unit of analysis
4. Determining the logic to link the data to the propositions	- Relying on the propositions as the logic to link the data to the propositions - Choosing an analytic technique, cross-case analysis
5. Determining the criteria to be used in data interpretation	- Doing the same things in the fourth suggestion for the fifth suggestion

3.3. Research Setting

This study was made in a Turkish foundation university. The medium of instruction was English in the Faculties of Economics and Administrative Sciences, Engineering and Architecture in the university. The students of the faculties had to take a ten-month English preparation class to start studying in their departments. When the students became first grade, they took a four-hour academic English course. Academic English course was organized differently depending on the faculties' requirements. This course was designed as two hours English for specific purposes (ESP) and one hour English for academic purposes in the Faculty of Engineering and Architecture, while

ESP was the core of academic English course in the Faculty of Economics and Administrative Sciences. There were 50 students in each class of these faculties on average. However, the medium of instruction in the Faculties of Communication, Education, Medicine and Health Sciences was Turkish. When this study was conducted, there were two groups of students in these faculties. The first group took English preparation training, but the second group did not take because the university cancelled the obligation of taking English preparation training for the faculties in which the percentage of English as the medium of instruction was 30%. Therefore, these groups took two different academic English course. The four-hour academic English course was organized as an EAP course for the students in the Faculty of Education and as ESP course for the students in the other faculties. The second group of students studied academic English classes as an English for general purposes course. The number of the students was almost 50 in each class of these faculties. There were ten EFL instructors working in the academic English department. Each instructor had to teach English in different faculties.

3.4. Participants

A qualitative study deals with describing, understanding and clarifying a human experience, so it can explain or describe the different aspects of the experience that make it idiosyncratic (Creswell, 2007; Dörnyei, 2011). Therefore, it is necessary to find the participants who can provide a researcher with rich and varied insights about the experience under investigation, so the researcher can maximize what he can learn from the experience. Consequently, the qualitative study requires purposeful sampling. This type of sampling enables the researcher to make his sampling line up with the purposes of the study (Creswell, 2007; Dörnyei, 2011).

Purposeful sampling requires a small sample size because a small sample size can produce saturated and rich data from the participants in order to find out the subtle meanings in the phenomenon under investigation (Creswell, 2007). In addition, three different strategies can be used in purposeful sampling, one of which is criterion sampling in which participants are chosen according to the predetermined criteria (Dörnyei, 2011).

Considering these issues, purposeful sampling was employed in this case study to describe, understand and clarify how different EFL instructors implemented language assessment literacy in their English classes in a Turkish higher education context. As a purposeful sampling strategy, criterion sampling strategy was used for this purpose. The criteria below were prepared and the participants were selected accordingly:

1. A participant should be autonomous in his/her language testing and assessment practices. That is, he/she can
 - a. choose his/her own type of assessment,
 - b. prepare his/her own tools for the type of assessment he/she chooses,
 - c. administer his/her own assessments, score them and interpret their assessment results,
 - d. develop valid grading procedures using his/her students' assessments,
 - e. announce his/her students' assessment results to different stakeholders,
 - f. make decisions about his/her instruction, students and class according to the assessment results and
 - g. recognize and take necessary precautions against any illegal and unethical testing and assessment practice.

Twelve EFL instructors worked at the Academic English department of the university and were autonomous as they were in charge of every step of language assessment and evaluation from choosing a type of assessment to using test results and to make decisions. Therefore, they met the criteria developed by the researcher. When asked for participating the research, only eight of them (2 female and 6 male instructors) accepted to participate in the study. The participants were asked to decide nick-names for themselves to be used in data collection, analysis and interpretation. They wanted to be mentioned with the names in Table 3.2 which gives the demographic information for each participant.

Table 3.2.

Demographic Information about Each Participant

Participant	Gender	Age	Experience	BA	MA	Weekly Teaching Hours	Number of Students
Deniz	Female	28	5	EL*	ETI*	21	More than 200 students
İlkbahar	Female	35	10	ACL*	ES*	21	More than 300 students
Aziz	Male	30	7	ELL*	ELL	27	More than 200 students
Beşiktaşlı	Male	28	7	ELT*	ELT	21	More than 200
Black Eagle	Male	28	5	ELT	ELT	24	More than 200 students
Crazy Soul	Male	35	7	ELT	ES	15	More than 200 students
Tahiri	Male	30	9	ELT	-	21	More than 250
Tiger	Male	30	9	ELL	ELT	24	More than 300 students

Note: * EL: English linguistics, ETI: English translation and interpretation, ACL: American culture and literature, ELL: English language and literature, ES: Educational sciences and ELT: English language teaching.

As Table 3.2 shows, the participants were between 28 and 35 years old and had between five- and ten-year teaching experience. They graduated from the different departments of the Turkish universities like ELT and ELL. They taught more than 20 hours every week and more than 200 students in their English classes.

3.5. Data Collection Tools

Interview, think-aloud protocol, observation, focus group discussion and document analysis were used to collect data. Details were given under five sub-headings.

3.5.1. Semi-structured individual interviews

The researcher used semi-structured interviews in order to obtain rich details and in-depth information (Rubin & Rubin, 2005; Seidman, 2006; Turner, III, 2010) about how each participant implemented language assessment literacy and its sub-components in their English classes. As a result of a comprehensive literature review, an overview of the research situation was obtained and broad questions were prepared to enable each participant to elaborate and detail their answers in an exploratory manner (Dörnyei, 2011). Figure 3.1 briefly indicates the development of each interview.



Figure 3.1. The development of each semi-structured interview

As Figure 3.1 indicates, the researcher first developed the content of the interviews. Each interview had open-ended, big, neutral, certain and expansive questions to create a good rapport and comfortable interview atmosphere (Rubin & Rubin, 2005; Seidman, 2006). The interviews started with the basic questions to set the tone and create the initial rapport with the interviewees, went on with the content questions supported with probes and ended up with a final closing question to enable the interviewees to finish the interview by ordering the questions from the easy ones to the difficult ones and by categorizing them depending on the topic areas (Dörnyei, 2011; Jacob & Furgerson, 2012; Turner, III, 2010).

Three interview protocols were prepared based on Jacob and Furgerson (2012) for three interviews after the interview questions were prepared. The protocols included scripts which gave information about (a) the research in terms of its aim, goal and conduction, (b) confidentiality, (c) consent, (d) the contact information of the researcher, (e) duration (f) a thanking statement and (g) the reminder of the contact issues. The protocols also emphasized the appreciation and value of responses (Colker, n.d.).

The interview questions and protocols were piloted through a systematic approach: editing, early pilot and full pilot (Rubin & Rubin, 2005; Seidman, 2006). Each interview was edited after three EFL instructors read and gave feedback about their wording, so their wording was improved. Second, they were practiced with two EFL instructors, so the researcher could check whether they could be understood well and whether the interview questions were in order logically. Finally, he did full piloting. The interviews were made with the three EFL instructors. Their timing and administration conditions were checked and the necessary changes were made.

The semi-structured interviews were made in Turkish and in three sessions to obtain sufficient in-depth, broad and rich data about the participants (Dörnyei, 2011). In addition, the first interview lasted between 45 and 60 minutes, was made face-to-face and was related to the participants' background information. The second interview was about the implementation of the seven sub-components of language assessment literacy. It lasted between 60 and 80 minutes and was made face-to-face. The third interview was made face to face or online in order to enable each participant to check the transcriptions and initial analyses of the data collected. It lasted almost 30 minutes.

3.5.2. Think-aloud Protocol

The researcher used think-aloud protocol to describe the cognitive process(es) which each participant used while preparing their questions because it helps to understand and describe what is focused on and how this information is structured during a task (Fonteyn, Kuipers, & Grobe, 1993). Therefore, inferences could be made about the process(es) each participant used while preparing their exams. Concurrent think-aloud protocol was used to have the direct verbalization of the cognitive processes with retrospective think-aloud protocol which the researcher used as a follow-up to have a broad picture of the cognitive process(es) of each participant (Fonteyn et al., 1993).

The five stages proposed by van Someren, Barnard and Sandberg (1994) were used in think-aloud protocol: setting, instructions, warming up, the behavior of the examiner and prompting and recording. At the determined times, each participant was visited and informed about what they should do and what think-aloud protocol was in their offices or flats. Then, they practiced think-aloud protocol, so they became them

familiar with the do's and don'ts of the procedure. Each participant was audio recorded. The researcher only interrupted them when they forgot to talk their thoughts aloud. Think-aloud protocol lasted between 30 and 135 minutes.

3.5.3. Focus group discussion

Focus group discussion benefits from group dynamics because they encourage participants to talk to each other, so participants explore and clarify their ideas, feelings and opinions less accessible in the individual interviews (Kitzinger, 1995). Therefore, a focus group discussion was made to verify the data collected in the individual interviews. The researcher prepared his focus group discussion questions from his research questions by following the suggestions made by Yıldırım and Şimşek (2013) such as being easy to understand for the participants, only related to one aspect of the phenomenon and presenting with a clearly prepared instruction.

The focus group discussion questions were ordered according to the dimensions formed in the individual interviews. Therefore, the focus group discussion started with the participants' perceptions about education, assessment and evaluation and the factors affecting their assessment and evaluation and ended up with the implementation of seven sub-components of language assessment literacy.

Seven of the participants joined the focus group discussion. Before the discussion was made, the researcher arranged a meeting room where everybody could sit comfortably, feel relaxed and discuss the questions without being disturbed by any noise. A focus group discussion protocol was prepared by using the same procedure of preparing the semi-structured interview protocols. The discussion was audio recorded. During the discussion, the researcher encouraged discussion, became a good listener and only interrupted the discussion when it lost its focus. The focus group discussion lasted almost 75 minutes.

3.5.4. Semi-structured, field and non-participant observations

The researcher used observation in order to reach first-hand data about the participants and to support the data he obtained from the individual interviews because observation enables an observer to describe and explain human behaviors holistically in

the behaviors' natural environment (Yıldırım & Şimşek, 2013). He used four semi-structured, field and non-participant observations in which he observed each participant in the participant's natural environment as an outsider and used a protocol to make an observation (Yıldırım & Şimşek, 2013).

Four different observation protocols were prepared based on Yıldırım and Şimşek's suggestions. The researcher first determined the dimensions of each observation in order to observe the phenomenon under investigation thoroughly in preparing his observation protocols. Second, he formed a part to describe the physical environments in which the observations were made. Then, he prepared three different parts to observe the social dimension of, the actions happening in and the language established in the field.

During observations, the researcher recorded everything related to the research purpose when they happened without making any subjective judgment. Observation notes were kept as descriptive and detailed as possible. He mentioned which notes reflect his personal judgments. The first and second observations were about two meetings made in order to choose their assessment methods and develop their grading system. They lasted 60 minutes and 30 minutes in order. The third and fourth observations were made to observe each participant in administering their exams, grading them and announcing their results. The third observations took between 30 and 135 minutes, while the fourth observations lasted between 30 minutes and 90 minutes.

3.5.5. Document analysis

The researcher used document analysis to triangulate the data because it provides supplementary research data which can be analyzed to verify the findings from the other sources (Bowen, 2009). Before analyzing the documents, he followed the five stages stated in Yıldırım and Şimşek (2013) and this process was shown in Figure 3.2.

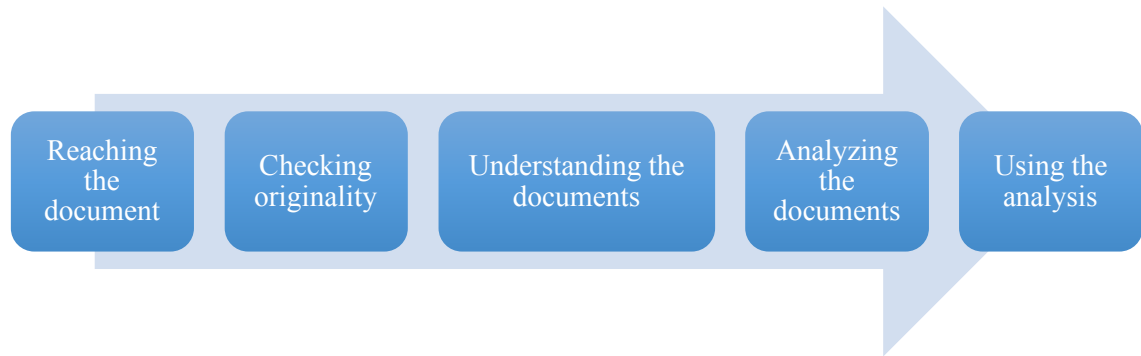


Figure 3.2. The processes of document analysis

As Figure 3.2 indicates, document analysis started with reaching the documents and deciding whether document analysis was needed. In the first stage, it was determined whether document analysis was needed, why it was needed, what kind of documents were needed and where and from whom documents could be reached. In this stage, the researcher decided to use the participants' quizzes, midterms, final exams and course books for his research purposes. For the second stage, the documents were collected from their owners. As the documents were taken from their owners, they were original and related to each participant.

For the third stage, Bowen mentioned that understanding the data is related to the purpose of using document analysis. As the purpose for using the document analysis was to support the data obtained from the other sources, the codes developed during the analyses of the observations, think-aloud protocol, focus group discussion and individual interviews were used to understand the documents.

In the fourth stage, the data were analyzed through a document analysis protocol which was prepared based on the five basic principles of language assessment. These basic principles are validity, reliability, washback, practicality and authenticity (Brown, 2004).

In the fifth stage, the researcher got the consent of each participant to use the results of the document analysis. The participants were ensured that they would not be damaged. In addition, the initial findings were shared with each participant before being reported. Some suggestions like balancing sensitivity and objectivity and deciding the sampling size according to the use of the document analysis were used in evaluating the documents (Bowen, 2009).

3.5.6. The relationships between data collection tools and sub-components of language assessment literacy

The second individual interview was the main source of data collection. The data collected from the second interviews were supported by the data obtained from the first interviews, observations, think-aloud protocol, document analysis and focus group discussion. The contributions of each data collection to understanding how each sub-component of language assessment literacy was implemented were shown in Table 3.3.

Table 3.3.

The Relationship between Data Collection Tools and Sub-components of Language Assessment Literacy

Sub-Component of Language Assessment Literacy	The Tool(s) Used for Collecting Data Related to the Sub-Component
1. Choosing assessment methods	- The first interview - The second interview - The first observation - Focus group discussion
2. Developing assessment	- The first interview - The second interview - Document analysis - Focus group discussion - Think-aloud protocol
3. Administering exams, scoring them and interpreting their results	- The second interview - Think-aloud protocol - The third observation - The fourth observation - Document analysis - Focus group discussion
4. Using assessment data in making decisions about students, instruction and curriculum	- The second interview - Focus group discussion
5. Developing a valid grading system	- The second interview - Focus group discussion - The first observation - The second observation
6. Communicating assessment results	- The second interview - The fourth observation - Focus group discussion
7. Recognizing illegal and unethical assessment practices	- The second interview - Think-aloud protocol - Document analysis - Focus group discussion - The third observation - The fourth observation

3.6. Trustworthiness of the Study

The positivist quality criteria of validity and reliability in quantitative research cannot be applied to qualitative research because of the latter's interpretative nature (Pitney, 2004; Shenton, 2004). Therefore, the criteria were replaced by an alternative framework in qualitative research which is composed of credibility, transferability, dependability and confirmability, all which make a qualitative study trustworthy (Lincoln & Guba, 1985). Table 3.4 indicates which strategies were used to make this study credible, transferable, dependable and confirmable depending on this framework.

Table 3.4.

Strategies Used for Trustworthiness

The Criteria	Strategies Used
1. Credibility	<ul style="list-style-type: none"> - Early familiarity with the culture of the organization - Triangulation - Tactics to help ensure honesty in informants - Peer scrutiny of the research project - Information about the researcher - Member checks - Thick description - Examination of the previous research findings
2. Transferability	<ul style="list-style-type: none"> - Detailed, elaborated, rich and descriptive information about the phenomenon, participants and research context
3. Dependability	<ul style="list-style-type: none"> - The use of overlapping data collection tools - Data collection at different times - Detailing research processes and research design - Member checks - Triangulation - Research audit trail
4. Confirmability	<ul style="list-style-type: none"> - The researcher's beliefs and assumptions - Research audit trail

- 1. Credibility:** As Table 3.4 indicates, the first criterion is credibility which refers to whether the findings in the research show what is actually happening in the context and to whether a researcher finds out what he actually wants to learn (Lincoln & Guba, 1985). For the credibility, the researcher used eight strategies based on the literature (Lincoln & Guba, 1985; Pitney, 2004; Shenton, 2004) as Table 3.4 indicates.

- a. The researcher depended on his familiarity with the culture of the organization because he worked with the participants for at least two years at the university.
 - b. He triangulated his data by using five different data collection tools and cross-checked the findings to be sure that an accurate understanding of the research topic was obtained.
 - c. Several tactics like getting the consent of the participants, withdrawing from the study anytime and protecting the participants' privacy were used to ensure honesty in the participants.
 - d. Peer scrutiny was used in designing the research, developing data collection tools and analyzing the data because Lincoln and Guba (1985) thought that peer scrutiny can prevent a researcher from inhibiting his ability to view the research.
 - e. The participants checked and gave feedback about the transcriptions of the interviews, focus group discussion and think-aloud protocol as well as their analyses, so member checks enabled the researcher to be sure that his transcriptions reflected what his participants wanted to explain and that his inferences about the participants were verified by them.
 - f. The researcher described his findings thickly in the Findings Chapter because thick description helps to reflect the phenomenon under investigation and its context to a certain degree well, so readers can determine to what extent the findings reveal the truth about the phenomenon.
 - g. He made a comprehensive literature review and compared the findings of this study with the findings of the previous studies in order to assess the extent to which the new findings were congruent with the previous findings.
 - h. The researcher is familiar with qualitative research. He published several articles related to language assessment and evaluation. He also worked as a Testing Office member for almost three years and taught language assessment course in the ELT department of the university.
2. **Transferability:** Transferability helps to determine whether the findings of the present study can be related to the similar contexts (Lincoln & Guba, 1985). Therefore, it is important to give rich, detailed and descriptive information about

the phenomenon, participants and research context (Pitney, 2004; Shenton, 2004). For the transferability of this research, all steps were detailed and elaborated more thoroughly.

- 3. Dependability:** Dependability shows how reasonable the particular findings of a qualitative study are according to the collected data (Lincoln & Guba, 1985). The researcher collected data with overlapping data collection tools at different times to cross-check and cross-validate the data collected (Brown, 2005). He also gave detailed information about the research processes and research design to serve as a prototype model for readers to evaluate and assess to what extent the proper research practices are followed (Shenton, 2004). Member checks, triangulation and two detailed audit trails were also used to make this study dependable (Pitney, 2004).

 - a. The research audit trail:** An audit trail requires a researcher to document how the data analysis is completed and how theoretical, methodological and analytical choices are made, so other researchers can monitor the whole mental processes and control his research decisions (Carcary, 2009). For these purposes, the researcher prepared Chapters 1, 2 and 3 as the audit trail of this study which documents how the data analysis was completed and how theoretical, methodological and analytical choices were made.
- 4. Confirmability:** Confirmability shows to which extent the findings of a qualitative study are the results of the experiences and ideas of the participants under investigation, but not the results of the characteristics and preferences of a researcher (Lincoln & Guba, 1985). For the confirmability of this study, the researcher triangulated his data, noticed the shortcomings of his data collection tools and their possible effects and used several strategies to overcome the shortcomings (Shenton, 2004). The researcher's beliefs and assumptions about the case under investigation were also mentioned for the confirmability of the research (Bugel, 2011; Lincoln & Guba, 1985) through bracketing and decentering which allowed the researcher to be open to the experiences and perceptions of the participants (Bugel, 2011). In addition, he prepared an audit trail to make the study confirmable.

- a. The researcher's beliefs and assumptions about the case under investigation:** The researcher prepared questions for four skills, grammar and vocabulary when he worked as a Testing Office member. He improved his assessment knowledge and practices by writing his questions and through peer feedback and self-assessment. In writing multiple-choice questions, he used his pre-service assessment training, but he learned how to administer, score and interpret assessments on the job. In academic English department, he chose assessment methods for different courses and developed his grading systems depending on the content of the courses and classroom activities. In ELT department, he used available grading systems for some courses, selected different assessment methods and developed grading systems for other courses in collaboration with the head of the department. He used four types of assessment methods, but he learned how to assess with performance assessment and personal communication on the job. He encountered different problems like not following standard grading procedure in grading writing and speaking exams during six years, but did not know the reason(s) of them.

3.7. Data Collection Procedure

The research was carried out in a Turkish university located in the Southern Anatolia in 2015-16 education year. The researcher first got a legal permission was from the university. Then, he made a literature review of assessment and language assessment literacies. After he finished the literature review, he prepared his first semi-structured interview questions based on the literature review. He piloted his first interview questions and then made the first interviews with the participants. Meanwhile, he made his first and second observations in which the participants chose their assessment method and developed their grading system. He followed the same preparation procedure with the second interviews and conducted it with the participants. For each interview, he first took appointments from each participant and made the interviews in each participant's offices at the scheduled times. Then, he prepared think-aloud protocols and took appointments from each participant. He informed each

participant about what to do, helped them practice the procedure and did think-aloud protocol with them in their offices or flats at the planned times before the midterm exams. In the midterm week, he observed how each participant administered their exams, scored their exams and announced their grades. After the final week, he and seven of the participants joined a focus group discussion in which the main concepts in the first and second interviews were discussed. Following the focus group discussion, he took a sample of the midterm, final exam and quiz from each participant and borrowed their course books for document analysis. Meanwhile, he transcribed the first and second interviews, think-aloud protocol and focus group discussion. Then, he content-analyzed them and document analyzed their exams and course books. Next, he interviewed each participant face to face or online third time, shared his findings with them and discussed the findings in the third interview. Figure 3.3 summarizes this procedure below.

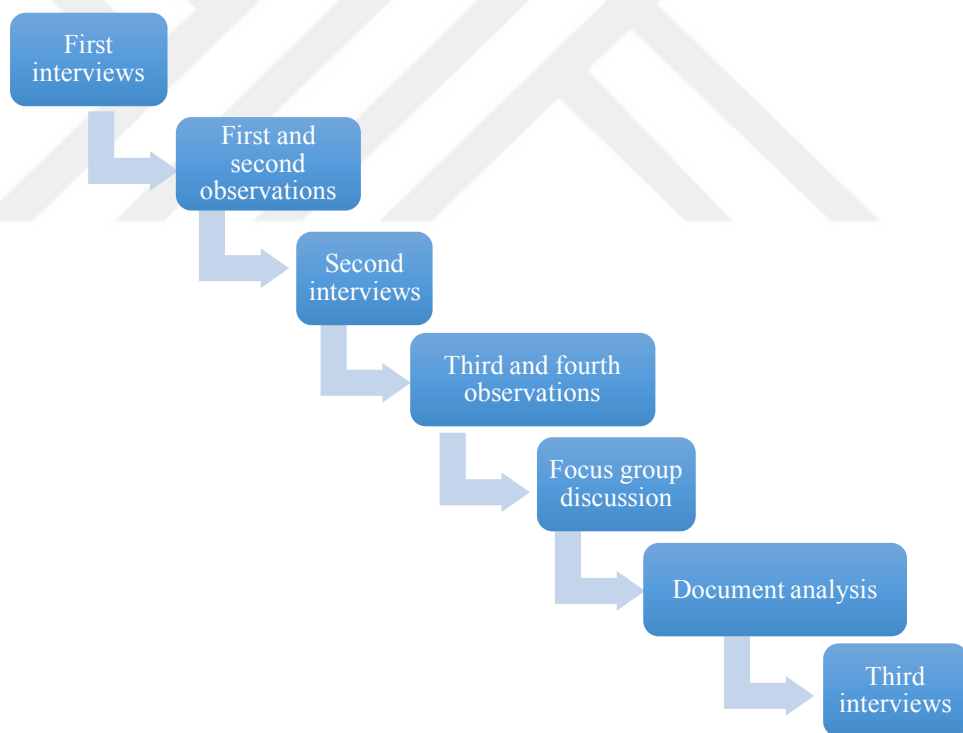


Figure 3.3. The summary of the data collection procedure

3.8. Data Analysis

The researcher used content analysis to reveal the concepts and relationships through four ways (Yıldırım & Şimşek, 2013). Figure 3.4 shows the process of content analysis.

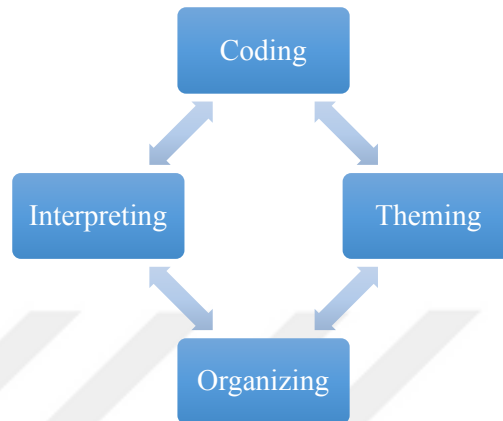


Figure 3.4. The process of content analysis

The researcher categorized the data into meaningful units and conceptualized what these meaningful units were by giving codes which explained the relationships in each meaningful unit. He read the data many times to code and used the codes derived from the data to name them. After preparing a code list, he found the themes which covered the codes in the list by finding out the similarities and differences among the codes, so he categorized the codes by placing the similar ones into a theme and explained the relationships among them. As a result of coding and theming, he had a system to organize the data. He organized and described the data with the quotations taken from the data collection tools according to this system in a way that his readers could understand. He presented the data collected from each different data collection tool by relating them to each other without adding his comments or interpretations to the analysis. He interpreted the data without conflicting with the description of the data in the end. He made explanations in order to make his data meaningful, to make logical conclusions from the findings, to reveal reason and result relationship and to show the importance of the findings.

He did these things to analyze the data he collected through the individual interviews and focus group discussion. He benefitted from the data obtained from the

second interviews for developing the codes to content-analyze observations and think-aloud protocol. He determined the codes to document analyze each participant's exams and course books according to Brown's (2004) suggestions to evaluate classroom-based language assessments.

3.9. Ethical Procedures

In this research, the researcher took a legal permission from the university in order to start his research because procedural ethics requires taking a permission from an institution to do research (Yıldırım & Şimşek, 2013). He got the consent of each participant at the beginning of the study and every time when he collected data from them. He kept his research purpose and procedure transparent to each participant, so they could ask their questions whenever they wanted. He always informed them about the stages of the study. He kept his collected data and audio recordings safe and confidential. He did not mention the participants' real names in data analysis and reporting because of confidentiality and respect to their private lives. As they were informed about every stage of the research and could easily monitor the research, he did not deceive them. He focused his findings and discussions on the data collected and did not influence any participant in data collection.

3.10. Conclusion

This chapter has given detailed information about the research method and design, data collection tools, the research's trustworthiness, data collection, data analysis and ethical procedures followed in the study. The next chapter will present the findings related to the implementation of the seven sub-components of language assessment literacy.

CHAPTER FOUR

4. FINDINGS

4.1. Introduction

This chapter explains the similarities and differences among the eight cases depending on the data collected from the data collection tools. It first mentions the analysis of the background information of each participant. Then, it presents the analysis related to the implementation of language assessment literacy in English classes.

4.2. Understanding Education and Teaching Approaches, Assessment and Evaluation Approaches and the Factors Affecting Assessment and Evaluation

This sub-heading represents the analysis of the first interviews based on the themes in Table 4.1.

Table 4.1.

Themes and Codes of the First Interviews

-
1. Education and teaching approach
 - 1.1. Definitions of education and teaching
 - 1.2. Teaching approach
 2. Assessment and evaluation approach
 - 2.1. Definitions of assessment and evaluation
 - 2.2. Associations related to assessment and evaluation
 - 2.3. Feelings related to assessment and evaluation
 - 2.4. Effects of assessment and evaluation on teaching practices
 3. Factors that affect assessment and evaluation
 - 3.1. Types of assessment used to assess
 - 3.2. The effects of different types of assessment on their education
 - 3.3. Pre-service assessment and evaluation training
 - 3.4. In-service assessment and evaluation training
 - 3.5. Self-improvement in assessment and evaluation
 - 3.6. Changes in assessment and evaluation approaches
 - 3.7. The purposes of using assessment and evaluation
 - 3.8. The things paid attention to in assessment and evaluation
 - 3.9. Difficulties encountered in assessment and evaluation
-

This sub-heading details the issues in Table 4.1.

4.2.1. The participants' education and teaching approaches

As Table 4.1 shows, the participants' education and teaching approaches were analyzed under two categories: their definitions of education and teaching and teaching approaches. This part presents the analysis in this order.

4.2.1.1. Definitions of education and teaching

As the quotation below indicates, Tahiri focused on the moral and materialist preparation of his students in his definition of education because he believed that education and teaching equip his students with necessary skills to deal with possible future problems in their personal and professional lives.

Tahiri (Interview 1[I1]): Education and training prepare an individual for the possible problems in the future. They are giving the material and emotional power to him to overcome the possible problems. If not, he will be in trouble. This must be the most important thing for an educator: preparing his students for the difficulties in future, for life and for their profession.

Similarly, Tiger concentrated on shaping his students according to what the society and institution want them to be as the quotation below shows.

Tiger (I1): Education and training are shaping. I can define shaping as turning the thing into what is desired. Namely, education and training are the shape the society or institution wants to be given to the raw material. What is important is giving a shape to the raw material.

Tahiri and Tiger focused on preparing students in terms of moral, materialism and society through education and teaching, but Aziz concentrated on preparing students for life by helping them endure pains and gain experiences through these pains as understood from the first interview excerpt below.

Aziz (I1): Sainthood is an important word for me. Like a saint who goes through life experiences and pains on the way to become a saint, a teacher can become a saint if he can stand the pain and improve himself. Therefore, education and teaching are pains for me and these pains enable me to improve myself and my students through experiences. Thus, education is an experience for me.

Like Aziz, Deniz focused on a different aspect of education and teaching. She thought that education and teaching prepare students for life through mutual exchange of information. To illustrate:

Deniz (I1): In my opinion, education and train are the share of information based on mutual interaction. That is, they are an exchange of information in terms of theory, ethics and practice.

In addition to preparing students for life in different aspects, İlkbahar thought education and teaching meet the needs of students. The following quotation clearly points out this issue.

İlkbahar (I1): If I want to define education and teaching, I can define them as increasing the learning outcomes of our students or others who need to learn new information by doing my best depending on their needs. ... because I care my students' taking pleasure and being happy while they are learning. For this, I think I sacrifice a lot.

Apart from Black Eagle, Crazy Soul and Beşiktaşlı, the rest of the participants defined education and teaching together. Their definitions focused on the fact that the individuals need to be educated and taught in order to be ready for dealing with what their future will bring to them in terms of ethics, morality, materialism, society, values and social interaction. This preparation process includes gaining experience, enduring pain, interacting mutually, shaping and meeting needs.

On the other hand, Beşiktaşlı believed education and teaching are two separate, but complementary concepts. He thought that education is a life-long learning process shaped by an individual's family, school and life experiences, while teaching is what students are taught at schools. To demonstrate:

Beşiktaşlı (I1): In my opinion, education is a process which a person uses to improve himself in every field and to be ready for life. His training, experience related to life in his family, interaction with his school friends, information, skills and experiences taught by his teachers are involved in this process. On the other hand, training is a more systematic, planned, scheduled process in which he is a student and taught depending on a target.

Like Beşiktaşlı, Crazy Soul made two different definitions of education and teaching. According to him, education is adopting the values of society and identifying one's own capacity, whereas teaching is what the school and life teach the individuals. The quotation below supports this finding.

Crazy Soul (I1): In fact, education and teaching are two different concepts. Education means the future of our [the society's] youth and country. Teaching includes specific things given at schools as the sub-branch of this education. Education exists at every stage of life. It is not for having a job, but for a person's understanding himself, preserving his values and realizing himself. Yet, teaching starts at primary school and finishes at the university. It is an institution which teaches a person a profession, makes him be specialized in that profession and helps him to have a job.

Black Eagle thought education is what is taught at school, but teaching is what the school and life teach the individuals. The following excerpt clearly indicates this issue.

Black Eagle (I1): Education and training are in fact complementing each other. I think one without the other one is not completed. Education is what is done at school. Training is what is done at school and in students' private lives and is also related to their social lives.

Beşiktaşlı, Black Eagle and Crazy Soul related their understanding of education and teaching to what is done at schools. Unlike the other participants who thought education and teaching should prepare students for life, they focused on the effect of life on education and/or teaching because they believed life, together with the school, teaches students.

4.2.1.2. Teaching approach

In terms of their teaching approaches, the participants preferred to teach eclectically. Eclectic method was used because of the following factors: (a) the number of their students, (b) their students' motivation and interest in learning English, (c) their course books, (d) their students' needs for learning English, (e) the content of their lessons, (f) their students' attitudes and behaviors toward their classes, (g) the topic taught, (h) their teaching environments, (i) the situations they were in and (j) their personal beliefs about teaching.

4.2.2. The participants' assessment and evaluation approaches

As Table 4.1 demonstrates, the participants' assessment and evaluation approaches were analyzed by focusing on their definitions of evaluation and assessment, associations and feelings that assessment and evaluation created for them

and the effects of assessment and evaluation on their instruction. This part represents this analysis by following the order in Table 4.1.

4.2.2.1. Definitions of assessment and evaluation

Beşiktaşlı told that assessment is calculating the result of what a teacher does and that evaluation is interpreting the results that the teacher obtains through assessment. On this issue;

Beşiktaşlı (I1): I think assessment and evaluation are related to each other, but different terms in nature. I define assessment as doing the calculation related to a process or a case. Evaluation is the process of giving meaning to the value that we [teachers] have as a result of the action we do and of the calculation we do. And the exams we make during the term are assessment tools. Our midterms, final exams and presentation are assessment tools, but the grades we have from them and their interpretations as pass and fail are our evaluation process.

Likewise, Black Eagle stated that assessment is obtaining a result for things like an exam and that evaluation is making a judgment depending on the result of assessment like pass or fail. The below quotation supports this finding.

Black Eagle (I1): Assessment can be an exam, a question, an activity that I expect my students to give an answer. Evaluation is a conclusion that I can make based on the assessment. How can we [teachers] make an evaluation? It is like you [a student] pass, fail or cannot answer. We can make such evaluations with exams.

Like them, İlkbahar defined assessment and evaluation as "... testing to what extent my students have learned what I have taught according to some criteria." Namely, according to her, assessment and evaluation are checking her students' learning.

Aziz believed that objectivity is very important for assessment and evaluation. Therefore, he believed that assessment is a tool used for providing objective results. These results are evaluated based on the grading system and reflect what a teacher thinks about his students. To indicate:

Aziz (I1): Because we [teachers] do this [evaluation] before entering the evaluation criteria about a person: You [teachers] form your certain prejudice without using numbers or assess the criteria in order to make this objective and scientific. For example, exams... When we assess our students through exams, the numerical values about them provide us with the scientific proof. Therefore, our interpretation method is an evaluation resulting from our grading system. It reveals our opinions about a student.

The explanations above show that half of the participants made different definitions for assessment and evaluation. Black Eagle, Beşiktaşlı, İlkbahar and Aziz considered assessment as checking what their students learn by calculating their progress in mathematical forms and evaluation as giving meaning to their students' assessment results for making conclusions about their students.

On the other hand, the rest of the participants made a single explanation for assessment and evaluation. According to Tahiri, assessment and evaluation "are measuring a student's learning and his teacher's achievement in teaching his courses." That is, he defined assessment and evaluation as checking his students' learning and his teaching.

Like Tahiri, Deniz told that assessment and evaluation are providing a teacher with feedback about his students' learning and using this feedback for checking the success of his teaching. The excerpt is related to this result.

Deniz (I1): Assessment and evaluation are the way of receiving feedback from the information taught for a while. Because assessment makes me think that we [teachers] assess the information or receive feedback from the information based on a standard. I should assess it somehow and evaluate the information I have taught in order to determine whether I am successful.

In addition, Tiger mentioned that assessment and evaluation are that a teacher self-assesses his instruction in order to determine its success by looking at to what extent his students have learned. To demonstrate:

Tiger (I1): Assessment and evaluation are asking for the return of what you [teachers] teach according to a criterion. That is, they check whether students acquire what they learn. People name them as written and spoken examination. You examine whether what you want is achieved and what the weaknesses are if not. Assessment and evaluation are made in order to see these. They provide these.

Crazy Soul made a similar definition to the ones of Tiger, Deniz and Tahiri. According to him, assessment is a means through which results are evaluated in terms of a teacher's instruction and his students' learning. To illustrate:

Crazy Soul (I1): In fact, assessment is a tool and process. Evaluation is the result which shows where we [teachers] can reach. Evaluation is to see to what extent we have taught and to what extent our students have learned.

The second group of participants focused on the self-assessment function of assessment and evaluation in their definitions. Assessment results provided them with data to determine the effectiveness of their teaching depending on to what extent their students learned what they taught in their classes.

4.2.2.2. Associations related to assessment and evaluation

Tahiri said in the first interview that assessment and evaluation were associated with “Exams, oral exams, students’ anxiety and not being able to speak.” Similarly, İlkbahar associated them with negativeness because she said “Positive things were not associated with assessment and evaluation. Assessment and evaluation remind me of exam anxiety, the focus of the exam and exam result.” Like Tahiri and İlkbahar, Deniz related assessment and evaluation to test because she stated “Unfortunately, assessment and evaluation are associated with being mechanic for me and test only comes to my mind when I hear assessment and evaluation.” Black Eagle also uttered “Failure, pass – fail and grade” were the associations of assessment and evaluation. In addition, Beşiktaşlı associated them with “Nervousness. That is, a nervous process, a nervous situation.” since he told “When I think as if I was in assessment and evaluation process, I feel nervous.” Aziz also associated assessment and evaluation with anxiety and nervousness because he mentioned “Whenever I think of assessment and evaluation, an anxious atmosphere and a nervous situation emerge in my mind.” Crazy Soul had a negative association for assessment and evaluation, but he differed from these participants as he associated them with tiredness. He said “First, assessment is a tiring situation for me especially when I have to obtain a result on paper because of administering my exam. Second, the number of students... because of grading exams.” As the first interviews indicated, assessment and evaluation created negative associations for most of the participants. However, they only led to a positive association (seeing the result of an effort) to Tiger because he explained “Assessment and evaluation prove how mature a person becomes.”

The participants mentioned several reasons for the associations listed above: (a) their professions, (b) legal issues, (c) making decisions about their students, (d) the types of the exams, (e) the number of the students, (f) seeing the concrete results of teaching efforts and (g) previous assessment experiences.

4.2.2.3. Feelings related to assessment and evaluation

Tahiri stated in the first interview that he felt angry, tiring and happy when he assessed his students. He felt angry because he said “A student tries to cheat during an exam and he does not have self-confidence, or I ask a very simple question, but he answers it incorrectly.” He added he felt happy “When students think they really learn by looking at exam results, I think that I have achieved and I am useful.” Black Eagle enjoyed assessing and evaluating because he explained “I teach lots of students. Seeing that they get high grades on the exams and performed well makes me really happy.” Deniz took pleasure while preparing exams because she stated “If I produce something, I check what I know or I become more creative”, but she felt assessment and evaluation were painful in crowded classes since she said “I become mechanic and cannot know what kind of a person I am in crowded classes while I try to avoid cheating.” İlkbahar also got excited when she assessed her students as she told “I really want to know whether I have taught well and whether my students have learnt well and can show their learning on the exam.” Similarly, Tiger got excited because he expressed “I want to find out whether my instruction has worked, so I wait to see the result. This waiting makes me curious and curiosity leads to excitement.” On the other hand, Beşiktaşlı felt anxious because he mentioned “I approach each stage of assessment and evaluation very cautiously and I believe that this cautiousness leads to anxiety.” Crazy Soul felt exhausted because he said “I get exhausted when I administer my exams. I have to deal with students’ attitudes, find exam venues, arrange students’ seating and prepare different booklets.” Aziz felt that he had an obligation to assess and evaluate because he stated “Being a teacher requires assessing and evaluating students.” As understood from the first interviews, assessment and evaluation led to both positive and negative feelings among the participants.

The reasons for these feelings included (a) the continuous control of the whole assessment procedure, (b) cheating, (c) seeing that their students learnt something and that they achieved their goals and objectives as teachers, (d) writing and preparing exams, (e) the number of the students and (f) arranging the whole exam procedure.

4.2.2.4. Effects of assessment and evaluation on the participants' instruction

As the participants explained in the first interviews, assessment and evaluation affected their teaching in several ways. According to Tahiri, Deniz, Black Eagle, Tiger, Crazy Soul, Beşiktaşlı and İlkbahar, they used their assessment results to self-assess their instruction, goals, course objectives and assessment tools in order to find out and overcome the problem(s) like asking questions above their students' levels of English. As a result, they could decide whether they achieved their goals and objectives in their courses and increased their awareness in assessment and evaluation. To do these, Beşiktaşlı and Crazy Soul determined a psychological grade level based on their students' classroom performances. In addition, Tahiri and Deniz used their assessment data to find out their students' weaknesses and help them to overcome these weaknesses. Black Eagle and Crazy Soul believed that their assessment practices were parallel to their teaching activities, so assessment and evaluation directed their teaching and shaped their expectations. Their students' grades affected İlkbahar and Tahiri emotionally. Designing her own grading system and varying her assessment tools based on the content of her lesson made Deniz feel fair and comfortable. While assessment and evaluation helped Aziz to understand their importance in education, they caused İlkbahar to believe that education was more important than assessment and evaluation. Therefore, she paid more attention to the fact that her students should enjoy what they would learn and be happy for learning. In addition, understanding the importance of the standardization of assessment and evaluation practices was the other effect of assessment and evaluation on teaching for Crazy Soul because if the standardization did not exist, the number of students' complaints would increase and being fair in grading would not be achieved.

As the focus group discussion pointed out, the participants believed that there is a direct relationship between their assessment and teaching. Their assessment and evaluation practices reflect their education and teaching practices. For example, Beşiktaşlı mentioned, "If a teacher is process-oriented, his assessment practices are process-based, but if the teacher is product-oriented, his assessment practices become product-oriented." They accepted that they were product-oriented because of the numbers of their students, their workload, their students' levels of English and the lack

of time though they wanted to follow their students' learning processes. In addition, they educated and taught their students according to the way they assessed their students. They also found out their students' weaknesses in their exams and tried to improve their students' weaknesses by doing extra activities related to their weaknesses. They might reduce the number of the questions which their students were not good at through their students' previous exam results, so they might increase the number of the questions which their students were good at answering.

4.2.3. Factors affecting the participants' assessment and evaluation

This part is composed of nine sub-parts. Each of these parts explains the factors which affect the participants in assessing and evaluating their students.

4.2.3.1. Assessment methods having been used to assess the participants as students

As stated in the first interviews, all participants were assessed with multiple-choice and open-ended questions when they were students. In addition, paragraph and report writing (to assess Tahiri), oral exam (to assess Beşiktaşlı, Deniz, Tiger, İlkbahar and Crazy Soul), project (to assess Tahiri, Black Eagle and Crazy Soul), presentation (to assess Beşiktaşlı) and portfolio (to assess Black Eagle, Crazy Soul and Beşiktaşlı) were also used to assess one or some participants when they were students.

4.2.3.2. The effects of different types of assessment on the participants' education

The findings of the first interviews revealed that the assessment methods used to assess the participants when they were students affected their education in several ways. First of all, the methods caused the participants to have the belief that if an exam encourages a student to think, study and produce something by using what he has learned and promotes his self-confidence, it is the correct assessment tool. In addition, Tiger believed that the assessment methods used when he was a student enabled him to have a job, so this belief affected his decision about the correct assessment methods. Like him, Tahiri emphasized that the assessment methods used in his student-hood

helped him to improve his language skills and to become an effective teacher, so he believed that those methods were the correct assessment tools. Besides, the assessment methods used affected the way Deniz, İlkbahar and Aziz studied, so they had negative attitudes toward the assessment methods which caused them to lack self-confidence and self-assessment. Consequently, Aziz did not support using assessment as a punishment in his present classes. Moreover, Deniz formed the thought that lesson content should be the main criterion in choosing assessment methods and determining the weights of different sections in the exams.

4.2.3.3. Pre-service language assessment and evaluation training

As stated in the first interviews, Beşiktaşlı, Tahiri, Black Eagle and Crazy Soul took pre-service training in assessment and evaluation because of their departments (ELT). Beşiktaşlı told “it [pre-service assessment training] has an effect. Its most clear effect is on what I do in terms of assessment and evaluation”, but he could not remember any specific information related to assessment and evaluation because he said, “a lot of years have passed since I learned at university.” That is, pre-service assessment training influenced his assessment and evaluation practices, but he could not remember any theoretical information owing to time.

Like Beşiktaşlı, Black Eagle found his pre-service assessment training effective because he used what he learned in deciding what and how to assess and preparing and organizing his assessment. The quotation clearly indicates this finding.

Black Eagle (I1): Of course, it [pre-service assessment training] has. For example, when I prepare an exam for a student, no matter what it is, it helps me to determine the criteria like what I should pay attention to, what my goal is and what I should assess because I learned them in my pre-service training. Besides, I learned how to prepare questions, what I should pay attention to in preparing questions and how to format an exam paper. We [teachers] learned them in our pre-service training.

Crazy Soul also found his pre-service assessment training effective because he learned the ways to make his exams valid and reliable which he still used in assessment and evaluation like Beşiktaşlı and Black Eagle. The excerpt illustrates this issue.

Crazy Soul (I1): ... we [teachers] took an assessment and evaluation course. We prepared our tests and checked their validity and reliability. For example, what did I do? I chose a topic in English. I developed a test after I taught it. It [assessment training] has effects. For example, can my exam which has 50 questions cover what I have taught in my classes? I check it. How are the questions going to be scored? I prepare it [the exam]

accordingly. It [assessment training] also includes preparing questions suitable to the level of students. It [assessment training] affects a lot.

Unlike Black Eagle, Beşiktaşlı and Crazy Soul, Tahiri found his pre-service training ineffective because of being too theoretical and lack of practice. Black Eagle, Beşiktaşlı and Crazy Soul considered their pre-service training as effective because according to them, their pre-service training enabled them to (a) understand how to prepare exams by focusing on what to pay attention in preparing their exams, what the purpose of assessment was and how to prepare the layout of the exams, (b) adjust the difficulty levels of their questions according to the levels of their students, (c) develop some criteria to judge the quality of their exams and (d) make their exams valid and reliable.

On the other hand, Deniz, Tiger, Aziz and İlkbahar had to take one course in their pedagogical formation training in order to be an English language teacher because of their departments. According to Deniz, the course in her pedagogical formation training was very theoretical for her to understand and the course teacher's attitude toward the course was negative; therefore, it disinterested and unengaged her. To demonstrate:

Deniz (I1): What can we [teachers] attribute this [ineffective pedagogic formation training] to? In order to take the certificate of teaching, I was told to take some courses unfortunately in my pedagogical formation training. In my opinion, assessment and evaluation course was something that consisted of numerical things. I remember that I failed in this course because presentations were made and composed of theoretical knowledge and numerical values. The course teacher did not pay enough attention to our [student teachers'] learning. As a result, I was not interested and engaged in the course. I think it was not given enough importance.

Like Deniz, Tiger did not think that the course in his pedagogical formation training affected his assessment and evaluation practices because he told "the course was not taught seriously" in the first interview. In addition, Aziz believed that the course teacher did not care the course and give him and his friends opportunities to practice what they learned as understood from the excerpt below.

Aziz (I1): I want to say frankly that it [pedagogic formation assessment training] is a serious problem. That is, we [student teachers] knew that the teacher of assessment and evaluation course did not provide us with an opportunity to practice what we [student teachers] learned and did not pay enough attention to the course. Therefore, we should not think that we can expect a student to have

the expectation that what he has learned will be useful in an environment where the teacher does not give importance to assessment and evaluation. Therefore, I had trouble in this course.

These participants did not find the courses they received in their pedagogical formation training effective because they were too statistical and lacked practice and because the teachers of the courses did not teach the courses seriously and give enough importance to them. The participants shared the same ideas about the assessment and evaluation course in their pedagogical formation training in the focus group discussion.

4.2.3.4. In-service language assessment and evaluation training

As the first interviews showed, only Beşiktaşlı, Deniz, Crazy Soul and Aziz took in-service training about assessment and evaluation. Aziz and Crazy Soul found the training effective. Aziz told that he understood how different types of assessment methods required students to use their different capacities, but Crazy Soul could not remember a specific effect like Aziz. However, Beşiktaşlı and Deniz found the training ineffective. Beşiktaşlı thought so because he could not remember anything related to the training.

4.2.3.5. Self-improvement in language assessment and evaluation

As Beşiktaşlı said in the quotation below from the first interview, he interacted and collaborated with his colleagues in assessing his students through peer assessment and feedback in terms of different issues like the level and quality of the exams, which he believed improved himself in assessment and evaluation.

Beşiktaşlı (I1): In fact, I did not attend any seminar related to, take any course about and read any book and article about assessment and evaluation, but I have always made exams with my colleagues during seven years. We [teachers] have made discussions and exchanged our ideas about the difficulty level of our exams, the averages of our classes and the quality of our exam questions. On the other hand, I did not do anything willingly and consciously to improve myself in assessment and evaluation.

Similarly, Tahiri shared his questions with his colleagues. He and his colleagues gave and received feedback to and from each other, which he thought helped to improve his assessment and evaluation practices. To indicate:

Tahiri (I1): Now, I give my questions to my colleagues in my department. They check my questions and give feedback about them in terms of their difficulty level and relatedness to the content coverage of the exam. This also contributes to my improvement.

Like them, Black Eagle had a chance to practice what he learned in his pre-service training by working in the Testing Office of the school. While working there, he and his colleagues checked their exams, gave and received feedback to and from each other about their exams and made necessary changes in their exams according to their peer feedback. Their feedback was on validity and reliability. The quotation below clearly points out this.

Black Eagle (I1): ... but I worked in the Testing Office for almost one year. I can say that I could learn see my mistakes thanks to my colleagues. For example, if I made a mistake, it was corrected by someone and someone else's mistake was corrected by me. We [Testing Office members] were not officially trained in testing and assessment, but we learned from each other. For example, there were assessment and evaluation criteria. What do we assess? Think that we prepare a question. Though it aims to measure X and I believe it does, but someone else in the office says it does not measure X and he cannot understand what it actually measures. The feedback is also on whether the question is clear and understandable and measures the correct thing.

In addition, Tiger observed his colleagues when they prepared their exams. He evaluated his observations and decided to use some of the things his colleagues did depending on his observations. To show:

Tiger (I1): I have completely done everything depending on my observation and experiences related to choosing assessment tools. I have improved myself in an old-schooled way in terms of education and training. That is, I observed my colleagues in terms of what they did. By looking at them, I developed my own way of assessment and evaluation. For example, you [a teacher] wonder something when your friends prepare their exam, so you observe what and how they do. You like some ways. They may do something different in their exams. For example, I did not use matching a lot in my exams, but when I saw that my friends used matching, I started to use it.

Apart from pre-service and in-service training, half of the participants improved themselves in assessment through peer assessment. They gave and received feedback to and from their peers while preparing their exams. In addition, one of them made observations when his peers prepared their exams, which he believed was a type of peer interaction and feedback.

Tahiri said “I follow what my high school teachers did, so I do what I observed in assessing my students,” and Tiger told “I also thought about how the exams were prepared by our teachers in the past. I use them in my assessment as I think they are the correct assessment tools” in the first interviews. As these excerpts reveal, some participants including Tahiri, Tiger, Deniz and Black Eagle improved themselves by transferring their assessment and evaluation experiences they gained when they were students to their present teaching contexts. They especially benefitted from this experience in the first years of their teaching careers.

Black Eagle stated “In fact, I had a chance to use the theoretical information that I learned at school in a place for almost one year” and İlkbahar explained “I attended a CELTA program abroad. I improved myself by using what I learned there, practicing my learning and cooperating with my colleagues” in the first interviews. As these quotations clearly indicate, four of the participants including Black Eagle, Crazy Soul, Beşiktaşlı and İlkbahar integrated their theoretical knowledge with their assessment and evaluation practices.

As the quotation below demonstrates, Crazy Soul improved himself in assessment and evaluation by gaining experience. That is, he adapted his pre-service assessment training and the knowledge he learned from his colleagues and from the Internet to his teaching contexts and conditions, which he believed improved his assessment and evaluation.

Crazy Soul (I1): In fact, a teacher learns how to teach and how to assess while he is working like the others who learn their jobs when they work. In fact, a teacher develops his own way depending on his teaching conditions by adapting what was told in his university courses and in-service workshops. Namely, he forms his own way of teaching by keeping what he has learned in mind and combining them with his teaching conditions. In this aspect, I improve myself by adding what I have learned from my colleagues and from the Internet to what I have already known.

Similarly, Aziz improved his assessment and evaluation practices by gaining experience through assessing and evaluating his students. The excerpt below points out this.

Aziz (I1): I improved myself especially by experiencing in assessment and evaluation. As you know, the most troublesome part at some universities in Turkey is the theoretical part. That is, they [universities] teach theories to

students, but do not give them enough opportunities to practice. The most troublesome thing for us [teachers] is that we know the theory, but cannot put it into practice. Therefore, if you ask a student in any discipline, he says “I did not learn anything at the university. I learned something in my working life. Similarly, I learned thanks to my experiences in my working life.

As the example quotations above illustrate, the participants gained experience by making assessment and evaluation on their own. While gaining experience, they improved their assessment knowledge and awareness in practice, self-assessed their experiences, transferred and adapted what they learned to their new teaching contexts. According to them, this was the most important and effective way of improving themselves in assessment and evaluation.

In addition to these ways, some participants followed some other ways to improve themselves. To exemplify, self-interest in assessment and evaluation and CELTA training for İlkbahar, writing a master thesis on assessment and evaluation for Black Eagle and studying the Public Personnel Selection Exam for Aziz and Beşiktaşlı improved them in assessment and evaluation.

In addition to the findings of the first interviews, the focus group discussion showed that three participants (Beşiktaşlı, Black Eagle and Crazy Soul) integrated their pre-service assessment knowledge with their assessment and evaluation practices, while the others improved themselves in assessment and evaluation by assessing and evaluating their students. The second group of the participants admitted that they did not know the theoretical aspects of what they did in assessing and evaluating their students. The participants accepted the contribution of peer interaction to the development of their assessment knowledge, but they were not sure about whether what they learned through this way was correct as understood from the focus group discussion.

4.2.3.6. Changes in assessment and evaluation approaches

As the first interviews demonstrated, the participants experienced individual changes in their assessment and evaluation approaches. Beşiktaşlı became more teacher-centered (grading what his students wrote) now though he was more student-centered (trying to give more grades to his students’ answers) in the first years of his career.

Tahiri changed the way of preparing his questions by asking questions about everything to asking questions about what he considered important for his students based on their ages, his experiences and his peer interactions). Deniz became more independent and creative, personalized assessment and evaluation process, adapted and linked her experiences with her teaching and realized different types of assessment methods like performance assessment now. Black Eagle prepared more student-centered exams and became more independent now though he used his previous experiences he had gained when he had been a student and prepared difficult exams like his teachers had done in the past. Tiger changed his style from using one type of question to using different types of questions. Crazy Soul benefitted from his own assessment knowledge more than he did in the past. He used his pre-service assessment knowledge and his previous experiences he had gained when he had been a student in assessing and evaluating his own students in the first years of his teaching career. Aziz became more idealistic now than he was in the past. He approached assessment and evaluation pragmatically when he started to teach, but he understood the long-lasting effect of assessment and evaluation on his students' lives, so his approach became more idealistic. İlkbahar was very strict in assessment and evaluation because she evaluated her students as successful or unsuccessful based on their grades in the first years of her teaching career, but now she was more flexible as she took into consideration her students' classroom performances.

4.2.3.7. The purposes of using assessment and evaluation

As understood from the first interviews, all participants used assessment and evaluation to check whether their students learned what they taught in their classes. Accordingly, Black Eagle and Crazy Soul wanted to give their students grades and Black Eagle also aimed at deciding who passed and failed. In addition, Beşiktaşlı's purpose was to make comments about his students and Deniz intended to create a link between her lesson and her students' participation. Apart from Beşiktaşlı and Black Eagle, the rest of the participants used assessment and evaluation to self-assess their instruction.

4.2.3.8. The issues paid attention in assessment and evaluation

As the first interviews indicated, Beşiktaşlı, Tahiri, Black Eagle, Crazy Soul and Aziz paid attention to the content validity of their exams in their assessment and evaluation. The students' levels of English were also important for Tahiri, Crazy Soul and Tiger in assessment and evaluation. Being fair in every step of assessment and evaluation was taken into consideration by Beşiktaşlı, Black Eagle and Tiger. Using different types of questions in their exams were essential for Deniz, İlkbahar and Aziz. Deniz also paid attention to avoiding inconsistency in her questions and receiving reliable feedback from her students; Black Eagle paid attention to objective grading and making his exams reliable; Aziz considered peer interaction, the options of the questions and the difficulty levels of the exams as important; Tiger cared asking questions which his students could understand; İlkbahar found asking questions related to what she considered important for her students essential in their assessment and evaluation practices. Besides, Crazy Soul paid attention to the weights given to different sections in the course book, Crazy Soul and Beşiktaşlı considered self-assessment important and Tahiri believed that teaching something to his students through his questions was important in his assessment and evaluation practices.

4.2.3.9. The difficulties encountered in assessment and evaluation

As the first interviews showed, all participants encountered some difficulties in their assessment and evaluation. Preparing content valid exams was exhausting, challenging and troublesome for Deniz, Beşiktaşlı and Black Eagle. The number of the students in their classes was also another difficulty for Beşiktaşlı, Crazy Soul, Tahiri and İlkbahar. Student behaviors in administering an exam was a challenge for Deniz and students' complaints about the exam results caused Deniz and Black Eagle to face some difficulties in their assessment and evaluation practices. Grading, fairness and objectivity were some of the other problems which Black Eagle, Tiger, Crazy Soul and Aziz had to deal with in assessing and evaluating their students. Being restricted to using one type of assessment methods because of the lack of time was an obstacle for İlkbahar to tackle in assessing and evaluating her students. Besides, Tahiri and Tiger encountered some difficulties in their assessment and evaluation practices due to the

type of the exam and believing that their assessment methods did not show their students' real performances. In addition, Aziz considered his students' taking the easy way out as a difficulty he had to deal with, while Crazy Soul had some concerns about whether his exams met the expectations. Beşiktaşlı was also concerned about the repeat students and he and Aziz had difficulty in self-assessing their instruction, goals and course objectives to find out and overcome any problem.

4.3. Implementing Language Assessment Literacy in Language Classes

This sub-heading mainly presents the analysis of the second interviews about how the participants implemented language assessment literacy (LAL) in their classes in accordance with the themes in Table 4.2.

Table 4.2.

Themes of the Second Interviews

1. Choosing assessment methods appropriate for instructional purposes
2. Developing assessments appropriate for instructional purposes
3. Administering exams, scoring them and interpreting their results
4. Using assessment results in deciding for student, instruction, school and curriculum
5. Developing valid grading procedures using students' assessment
6. Communicating assessment results to students and other stakeholders
7. Recognizing unethical, illegal and inappropriate assessment methods and uses of assessment information

The analysis of the second interviews was also supported by the analysis of the focus group discussion made based on the codes in Table 4.3.

Table 4.3.

Codes of Analyze the Focus Group Discussion

1. Choosing assessment methods
2. Developing assessments
3. Administering assessments
4. Scoring assessments
5. Interpreting assessment results
6. Using assessment results
7. Developing a grading system
8. Communicating assessment results
9. Recognizing unethical and illegal assessment practices

As Table 4.3 indicates, the codes are closely connected with the themes developed to analyze the second interviews. Therefore, the analysis of the focus group discussion was presented with the analysis of the second interviews in this part.

4.3.1. Choosing appropriate assessment methods for instructional purposes

This part gives information about how the participants chose their assessment methods relevant for their instructional purposes by focusing on the codes developed and used to analyze the first sub-component of LAL as Table 4.4 indicates.

Table 4.4.

Codes of the First Sub-component of LAL

1. Choosing assessment methods appropriate for instructional purposes
1.1. Definition of measurement error
1.1.1. Sources
1.1.2. Types
1.2. Definition of validity
1.3. Valid and invalid measurement data
1.4. The types of the assessment methods the participants used during their teaching career and for their present classes
1.5. Choosing assessment methods
1.5.1. Purposes
1.5.2. Strengths and weaknesses

The issues in Table 4.4. will be explained in detail in this part.

4.3.1.1. Definition of measurement error

Beşiktaşlı thought that measurement error is related to the content of the exams in terms of what the exam assessed and what it was supposed to assess. On this issue;

Beşiktaşlı (Interview 2[I2]): In my opinion, measurement error means an error about what we [teachers] want to assess because of some reasons. For example, we have an assessment tool and think that there is a problem with its content. Consequently, the result it provides is not what we want to assess, so it leads to an error. The problem with the assessment tool causes an error and avoids our reaching our goal. This is what measurement error is.

In addition, Tiger believed that not asking questions about what is taught to students and not adjusting the difficulty level of the questions depending on students' levels of English are measurement error. The quotation below points out this case.

Tiger (I2): Measurement error may be asking something that has not been taught. In my opinion, it is an error. That is, asking what has not been taught in class is an error. At the same time, I think preparing very easy or difficult questions is also an error.

Like Beşiktaşlı and Tiger, Crazy Soul thought that not asking questions from what students are taught is measurement error as well as ignoring their needs to the definition of measurement error. For instance;

Crazy Soul (I2): Measurement error? It may be asking questions that are not prepared depending on the students' needs or that are not related to what students have learned in their classes. It can be thought as asking questions from the fifteenth unit though the exam covers the units from one to ten.

Besides, İlkbahar thought that not asking questions from what students are taught and not preparing questions relevant to the students' levels of English are measurement error. The quotation below clearly supports this finding.

İlkbahar (I2): I can say that measurement error is the mismatch between what is aimed to measure and what the tool measures in this process. It is like preparing an exam not suitable for the students' levels or asking some questions from what is not taught in class.

Apart from relating measurement error to not asking questions from what students learn in class, to being unsuitable to students' levels of English and to not meeting students' needs, Tahiri believed that if a student's grade is very different from his real performance, this situation is called as measurement error. To show:

Tahiri (I2): Measurement error is the big difference between a student's exam grade and his effort in the class which deserves a higher grade. In other words, he is very hardworking, but his grade is low. Namely, it is the big difference between his real classroom performance and exam grade.

Like Tahiri, Black Eagle made a different definition. He stated that measurement error is "... making any change in a student's performance knowingly or unknowingly." In a different definition, Deniz stated that measurement error is "the problems we [teachers] have or the things which are ignored in our exams."

As understood from the participants' definitions of measurement error, if a teacher asks questions from what he does not teach in his classes and prepares questions which are not relevant to his students' levels of English and do not meet their needs, he causes measurement error. In addition, the difference between students' grades and real

performances, affecting their grades and the problems ignored in the exam questions are also named as measurement errors.

4.3.1.1.1. Sources of measurement error

In terms of the sources of measurement error, every participant thought that teacher was the primary source of errors as understood from the second interviews. Almost all participants except for İlkbahar and Deniz believed that exam was the secondary source. In addition, Aziz, Crazy Soul, Black Eagle and Deniz mentioned that students could also lead to measurement errors.

4.3.1.1.2. Types of measurement errors

As the second interviews pointed out, while Tiger, Deniz and Tahiri could not remember anything about the types of measurement errors, İlkbahar, Aziz, Crazy Soul, Black Eagle and Beşiktaşlı remembered either systematic error or random error, or both. They understood different things from systematic and random errors. İlkbahar and Crazy Soul thought that random error is making mistakes without noticing and unexpectedly though Aziz and Black Eagle believed that it means being subjective in grading because of a teacher's feelings about one specific student. Aziz and Black Eagle defined systematic error as giving or reducing the same points to or from the students. However, Aziz and Black Eagle were confused because they made the same definitions of standard and constant errors: giving or reducing the same points to or from students. Apart from them, Beşiktaşlı, İlkbahar and Crazy Soul though that systematic is making the same mistakes continuously in the exams.

4.3.1.2. Definition of validity

Tahiri said "Validity is probably asking questions related to what students learned in order to determine whether they learned" in the second interview. In a similar explanation, Black Eagle connected his understanding of validity with "... asking questions related to what we [teachers] teach. That is, the content of the exam matches the syllabus and curriculum of the course" in the second interview. Like them, Deniz told "I have taught the course for a while. If my assessment tool is consistent with what

I have done in my classes, it is valid.” That is, she believed that validity is creating a connection between what is done in a class and what is assessed in an exam. Crazy Soul’s definition of validity is similar to the ones above. He related his definition to the connection between an assessment tool and course objectives because according to him, the assessment tool indicates whether his students meet his expectations as the quotation below clearly points out.

Crazy Soul (I2): Validity of assessment? In fact, it is that an assessment tool covers what it is supposed to measure. That is, it can be thought of whether this assessment tool meets the objectives that students are supposed to achieve because validity is that the assessment tool provides what is expected of students.

Like Crazy Soul, Beşiktaşlı defined validity as “whether our [teachers’] assessment tool is suitable to its assessment purpose” because he believed that it reveals “whether our [teachers] assessment tool provides results related to what we want to achieve.”

In addition to the definition of validity as asking questions related to what is taught in class, Aziz also added preparing clear, understandable and readable questions with clear options to his definition of validity as seen in the quotation below.

Aziz (I2): Validity of assessment and evaluation... First of all, I can say that an exam must be relevant to the material, that is, the material we have really studied in our classes in order to be valid. The second thing is that questions must be written clearly and that the options should not be very close to each other. Therefore, validity should be determined depending on any physical errors like a spelling mistake or written in such a small font size that students cannot read and on being relevant to the in-class material. I think this makes an exam valid.

The others also made their own definitions of validity. Tiger believed that validity is “that an exam is suitable to the students’ levels”, while İlkbahar added, “using different types of questions and following the pre-determined criteria” to her definition of validity.

As understood from the participants’ explanations about validity, it is seen that most of them related validity to content validity. In addition, some participants added preparing questions suitable to students’ levels of English, avoiding any problem that affects students’ reading and understanding questions negatively and varying the types of questions in exams to their definitions of validity.

4.3.1.2.1. Types of validity

While İlkbahar and Tahiri could not remember anything about the types of validity, the others were familiar with content validity (asking questions related to what was taught in a class) as the second interviews indicated. Aziz and Deniz were familiar with face validity which is preparing readable, understandable and clear questions for Aziz and which is related to the form of the exam for Deniz. Aziz and Crazy Soul also knew criterion-referenced validity. Aziz believed that criterion-referenced validity is preparing realistic criteria to decide who is successful, whereas Crazy Soul believed it is assessing a teacher's goals and objectives. In addition, Deniz thought that construct validity is related to the form of an exam.

4.3.1.3. Valid and invalid assessment data

İlkbahar, Aziz, Deniz, Tahiri and Beşiktaşlı explained in the second interviews that if a teacher makes his exam content valid, the assessment data become valid. İlkbahar, Aziz and Tiger also thought that the data he obtains become valid if his exam is suitable for his students' levels of English. This finding also includes being suitable to their ages for Tiger and scoring according to the difficulty levels of questions for Aziz. Black Eagle, Deniz and Beşiktaşlı believed that if his exam is free of measurement errors, the data he obtains become valid, in addition. Besides these issues, writing conventions are a key indicator of valid assessment data for Aziz and reflecting a teacher's expectations from his students is what makes assessment data valid for Crazy Soul. In addition, Aziz claimed that if an exam is reliable, its data is also valid.

4.3.1.3.1. The effects of valid and invalid data on the participants' instruction

The findings of the second interviews indicated that invalid assessment data cause Black Eagle, Deniz, Tahiri and İlkbahar to self-assess their teaching and assessment practices in which they try to find out the reasons for invalid data and to make necessary changes in their teaching and assessment practices. Deniz also focuses on her students to understand who causes invalid data. If her students lead to invalid data, she wants them to change themselves. Black Eagle and Deniz mentioned that they

go on doing the same things in their teaching and assessment practices if their data are valid. On the other hand, Crazy Soul told that he tries to find out and overcome his weaknesses if his data are valid, but if not, he self-assesses his exams in terms of its content validity. Tiger believed that invalid assessment data cause a teacher to make wrong decisions about educating his students accordingly and so he cannot achieve his goals. Beşiktaşlı used his valid assessment data to evaluate his course, while Aziz related the effect(s) of valid and invalid assessment data to the concern about increasing workload because he believed that the teacher wants to reduce his workload, so he does not care about their effect(s) a lot.

4.3.1.4. The types of the assessment methods the participants used during their teaching career and for their present classes

The participants, except Black Eagle, used four types of assessment methods (selected response, constructed response, performance assessment and personal communication) in their previous classes as the second interviews revealed. Black Eagle used three types of assessment methods except for personal communication in his previous classes. Every participant chose selected response as their assessment method for formal assessment in their present classes. Tiger chose constructed response in addition to selected response for formal assessment and decided to use personal communication for informal assessment. Crazy Soul also chose performance assessment for formal assessment. Tahiri selected constructed response and personal communication for informal assessment in his classes. Deniz also selected performance assessment, constructed response and personal communication to assess and evaluate her students formally in her different courses.

4.3.1.5. Choosing assessment methods

The first issue that İlkbahar, Aziz, Crazy Soul, Black Eagle and Tahiri paid attention to in choosing their assessment method (selected response) was the high number of the students in their classes as told in the second interviews. The second important issue for İlkbahar, Aziz, Black Eagle, Tahiri and Beşiktaşlı in the selection process was reducing their workload because they thought that selected response is easy

to grade and administer. The third one that İlkbahar, Black Eagle, Deniz and Tahiri cared was the mutual decision about selected response, but Black Eagle said that the decision was under the effect of the previous course teachers' experiences with the use of different assessment methods in academic English classes. Beşiktaşlı and Aziz thought validity is essential in choosing an assessment method. Aziz also paid attention to measurement error and Beşiktaşlı took into account objective grading in selecting their assessment methods. İlkbahar decided to use personal communication for informal assessment in her classes because she thought that it enables her to check her students' learning depending on their participation. On the other hand, Crazy Soul and Tiger decided to use different assessment methods as well as selected response in their classes. Crazy Soul considered enabling his students to use English both inside and outside the class when he chose performance assessment depending on his experience and assessment knowledge. Tiger paid attention to the similarity between his classroom activities and exam questions, validity and deductions about the types of the questions he reached based on his observations when he chose constructed response as well as selected response. In addition, Deniz taught three different courses. She chose selected response to use in her academic English classes because of the mutual decision about it. If she had the previous course teachers' syllabi in her elective courses, she evaluated the assessment methods chosen by the previous teachers. If she did not find the grading systems in the syllabi logical, she changed them in interaction with her coordinator and previous course teachers. If she had to choose her own assessment methods, she paid attention to the name of her course, its content and its goals in selecting her assessment methods.

In addition, the focus group discussion showed that most of the participants thought their instructional decisions were idealistic, but the features of the educational system at their university limited them. According to İlkbahar and Beşiktaşlı, their weekly teaching hours, the number of the students in their classes and the syllabi affected the participants' decision-making process in choosing their assessment methods. Aziz also told that the participants made decisions which were idealistic, but they did not care them in practice because of the reasons İlkbahar and Beşiktaşlı explained for choosing a certain assessment method. In addition, Crazy Soul confessed that the participants did not pay a lot of attention to assessment and evaluation while

making instructional decisions about their course goals and objectives. Such decisions might change during the semester as Aziz told. Crazy Soul emphasized that their students' levels of English caused them to choose a certain assessment method. Deniz also added that practicing what was decided as instructional goals and objectives required some changes in assessment and evaluation during the term. On the other hand, Tiger thought that a teacher should take the initiative like him in choosing his assessment method because he believed that choosing an assessment method was his responsibility and was up to his wishes.

In addition, the participants made two meetings to choose their assessment method for the spring term of 2015-2016 education year and the researcher observed them in these meetings. In the first meeting, they wanted to go on using selected response in the second term because Crazy Soul, İlkbahar, Beşiktaşlı and Tahiri shared their previous experiences with the new members of the department (Deniz, Tiger, Black Eagle and Aziz). They told that they had wanted their students to prepare a presentation about a topic related to their majors, to write a report of their presentations, to make their presentations in their classes and to submit their reports on their presentation days. However, they mentioned that their students had copied and pasted information from Wikipedia without doing their own research. They had used this information in preparing their presentations and writing their reports without changing. They had not prepared their reports in the expected format and made their presentations on the scheduled days. Yet, they had complained about their grades when their grades had been announced. They had claimed that they had not deserved those grades because they had spent a lot of time on preparing their presentations. The old members of the department told that watching their students' presentations, giving grades to them, checking who had made the presentation on time, controlling who had submitted the report on time, reading their reports, grading the reports and dealing with the students' complaints about their grades had increased their workload. In addition, they added that their students had not followed the presentation rules though the participants had explained the rules to them many times in their classes. Moreover, some of the new participants told they wanted to give their students a teacher evaluation grade because some students were very good in their classes, but they got low grades from their exams and failed, so they wanted to help such students by giving an evaluation grade.

Nevertheless, the old members told that they had given such a grade to their students in the previous terms, but their students had complained about their grades by making such comments as “Why did you give this grade to me?”, “I came to all of your classes, but my evaluation grade is low, why?” and “Student X did not come to your class regularly, but you gave him a higher grade than me, why?”. Even though the students had been explained how that grade had been given to each student, they had not taken into consideration these explanations in their complaints. The old participants told their workload had increased owing to these complaints. As a result, the old participants had given up giving teacher evaluation grades. They also told the new members of the departments that selected response avoided such problems. It enabled them to grade their students’ papers objectively and give the grades which each of their students deserved depending on their knowledge. Considering these issues, all participants finally decided to use selected response as an assessment tool.

4.3.1.5.1. Purposes for choosing assessment methods

It was found in the second interviews that the main purpose of choosing assessment methods for all participants was to check their students’ learning. In addition, Beşiktaşlı chose selected response to decide who would pass and fail. Tiger also selected constructed response to check whether his students could use what they learned, while Tahiri chose it to assess his students informally in his classes. Crazy Soul selected performance assessment to help his students to use English and improve their research skills. Deniz chose different assessment methods because she wanted her students to understand the importance of the things studied in her classes and to produce something by using English. İlkbahar decided to use personal communication in her classes for engaging her students more, checking their understanding and following their participation, whereas Tahiri determined to use personal communication in order to make his students evaluate his classes at the end of the term.

4.3.1.5.2. The strengths and weaknesses of the chosen assessment methods

As the second interviews showed, every participant seemed to be familiar with the strengths and weaknesses of the methods they chose to assess their students in their classes. İlkbahar, Aziz, Crazy Soul, Tiger, Black Eagle and Deniz did not believe that

selected response assesses their students' real performance because the students do not produce anything in this way. İlkbahar, Crazy Soul and Tiger told that their students might answer the questions without knowing; therefore, Tiger and Black Eagle believed that it does not assess their students' real learning. According to Aziz, it shows what his students have learned in terms of knowledge. While Deniz believed that it may lead to exam anxiety, Beşiktaşlı thought that it enables him to grade his students objectively. One of its strengths was being easy to grade for İlkbahar, Crazy Soul, Black Eagle and Tahiri and being easy to administer for İlkbahar. Tahiri also mentioned that his students' familiarity with it is another strength of it, whereas İlkbahar told it is easy for her students to answer. Being time-consuming and difficult to prepare was its weakness for Beşiktaşlı and Tahiri. Tahiri also emphasized that the possibility of cheating in selected response is another weakness of it.

In terms of constructed response, Tiger and Tahiri thought that it shows their students' real performances and whether they can use what they have learned. However, Tahiri considered it as labor intensive, while Deniz thought that it leads to subjective grading and change in her students' study routines.

In terms of performance assessment, Crazy Soul believed that it may lead to subjective grading, but it enables his students to use English outside the class. Likewise, Deniz thought that it may lead to anxiety, but it enables her to double grade her students' presentations.

In terms of personal communication, creating a comfortable classroom atmosphere in which her students are very active and she can check their real learning is its strength for İlkbahar. In addition, it reveals his students' real feelings about hi and his teaching, which Tahiri thought is its strength.

4.3.2. Developing appropriate assessments for instructional purposes

This part informs about how the participants developed appropriate assessments for their instructional purposes by focusing on the codes developed and used to analyze the second sub-component of LAL as Table 4.5 indicates.

Table 4.5.

Codes of the Second Sub-component of LAL

2. Developing assessments appropriate for instructional purposes
2.1. Preparing exams in relation to the chosen assessment methods
2.2. Types of questions used in the chosen assessment methods
2.3. Providing validity

As Table 4.5 points out, this part first explains how each participant prepared their exams. It also presents the types of the questions the participants used in their exams and the ways they used to make their exams valid.

4.3.2.1. Preparing exams in relation to the chosen assessment methods

The findings of the second interviews demonstrated that the participants first checked what they taught in their classes before preparing their exam questions. While checking what they taught in their classes, Tahiri and Black Eagle focused on what they considered important for their students and eliminated the parts that they did not consider important. Crazy Soul added that he chose his exam listening audio and reading passages similar to the classroom ones in terms of their topics and lengths. The second issue that each participant took into account was their students' levels of English. For example, Aziz emphasized the importance of writing readable and clear questions and instruction that his students could understand. The third one was the types of the questions. They used the types of the questions that their students were familiar with because of the classroom and course book activities. Accordingly, İlkbahar tried to use different types of the questions because she believed that different types of the question activated her students' different capacities. Deniz, Crazy Soul and Tahiri used different types of assessment methods in their classes. Deniz chose her performance assessment based on her classroom activities: reading, discussing and presenting ideas in one of her courses. Besides, she chose her writing exam questions depending on her classroom activities: essay writing. Tahiri used constructed response as an informal assessment in his classes. He developed an in-class writing activity based on what he studied in his classes at that moment. In addition, Crazy Soul decided to use the presentation as a performance assessment because he taught how to make a presentation to his students in his classes. In writing their questions, Beşiktaşlı, Deniz and Aziz paid

attention to the options of their questions. While Beşiktaşlı thought that their lengths should be similar, Deniz and Aziz thought that two options must be very challenging, one must be less challenging and the last one must not be related to the questions' answers. In terms of writing their questions, Deniz told that she selected among the questions if she had a question pool. Tiger said that he developed his multiple-choice questions from his course book and course materials and used the questions he asked in his classes.

In addition to the second interviews, each participant joined a think-aloud protocol activity in which they prepared their midterm exams, through which the researcher aimed to find out the cognitive processes they used in writing their exams questions. The participants also gave samples of their quiz, midterm and final exams to the researcher, so the researcher analyzed the documents in order to check whether the participants followed what they told in the second interview and what they did in the think-aloud protocol.

4.3.2.1.1. The participants in action

The think-aloud protocol which aimed to reveal the cognitive processes the participants prepared their midterm exam questions was analyzed according to the codes in Table 4.6.

Table 4.6.

Codes of the Think-Aloud Protocols

-
1. Starting to prepare the exams
 2. Choosing reading passages, listening audio and/or words
 3. Deciding what to ask
 4. Preparing questions
 5. Self-assessing the written questions
 6. Evaluating the available questions
 7. Finalizing the preparation of the exams
-

This part explains each code by following the order indicated in Table 4.6. In addition, each participant was responsible for preparing their own midterm exams or some parts of the midterm exams since some participants were partners and were supposed to prepare the midterm exams together:

1. Beşiktaşlı, İlkbahar and Crazy Soul prepared listening, reading and vocabulary questions for their midterm exams. Tiger prepared listening, vocabulary and open-ended questions in his midterm exam.
2. Tahiri prepared vocabulary and grammar questions while his partner, Aziz prepared listening and reading questions for their midterm exams.
3. Black Eagle was responsible for preparing listening and vocabulary questions in his midterm exams, while his partner was supposed to prepare reading and grammar questions.
4. Deniz was in charge of preparing grammar, listening, vocabulary, pronunciation and reading questions in her midterm exam.

4.3.2.1.1.1. Starting to prepare the exams

Before starting to prepare their exam questions, Tahiri, Deniz, Black Eagle and Aziz first checked what they taught in their classes, while Crazy Soul, Beşiktaşlı, Tiger and İlkbahar secondly checked what was studied in their classes. Crazy Soul and Beşiktaşlı first decided how to prepare their exams, Tiger first thought about how to start his exam and İlkbahar first brainstormed about the structure of her exam. Every participant focused on how to start writing their exam questions as a first or second step. Black Eagle, İlkbahar and Aziz started preparing their exams with listening, Tiger and Crazy Soul began with vocabulary, Beşiktaşlı started with reading and Deniz and Tahiri began with grammar.

Each participant had to choose a starting point for themselves. It might be listening, reading, grammar, or vocabulary. Therefore, they were supposed to select a listening audio, reading passage and/or words to ask in their exams. They followed different ways to find out their reading passages and/or listening audio. Beşiktaşlı benefitted from the CD of his course book, Deniz from the test book and its CD of her course book, Tiger from Youtube, Aziz from the Internet, Crazy Soul from his previous exam and İlkbahar from the Internet and one of her colleagues in finding midterm reading passages and/or listening audio.

4.3.2.1.1.2. Choosing reading passages, listening audio and/or words

As the quotation from the think-aloud protocol indicates below, Beşiktaşlı paid attention to the level of an exam reading passage, the similarity of the topics between the classroom reading passages and the exam reading passage and the words used in the exam reading passage in choosing a reading passage for the midterm exam.

Beşiktaşlı (Think-aloud protocol [TAP]): I found a passage called neuro-marketing. Check whether its content was related to the students' department by reading fast. These topics and words are the topics and words that we [the teacher and his students] always talk about in our classes. It is related to the students' department. The words are similar to the ones that we have studied in our classes, but the language used is more difficult than the one used in the reading passages that we have studied. Maybe, I can simplify the sentences, so I can use it in the exam.

Similarly, these issues were important for Aziz, Crazy Soul, İlkbahar and Deniz. In addition, whether the reading passages gave a lot of opportunities to prepare questions was important for them. The length of the reading passage and the time necessary for their students to read it and answer its questions were taken into consideration by Crazy Soul, İlkbahar and Aziz.

When choosing a listening audio, Aziz, İlkbahar, Crazy Soul, Tiger, Beşiktaşlı and Deniz took into account their students' levels of English and the similarity between the topics of their chosen audio and the ones of the classroom listening audio. That the audio was understandable by their students and that it had a clear and audible recording were taken into consideration by these participants. Crazy Soul, Aziz and İlkbahar paid attention to the length of the audio. İlkbahar shortened the audio which she found for her midterm exam. İlkbahar, Beşiktaşlı and Tiger checked whether the audio included the words that they taught in their classes. In addition, having a clear instruction was important for Aziz and being able to prepare questions from the audio was essential for Beşiktaşlı in choosing their midterm audio. These participants preferred to choose their audio from the websites or CDs, but Black Eagle preferred to write his own script. He chose two of the topics which were studied in the listening parts of his classes and which he believed he could integrate with each other easily in writing his own script. He also paid attention to the points that the others cared in choosing their midterm audio from the websites and CDs. He enhanced his script while he was writing his questions

because he sometimes had difficulties in preparing a question from the script, so he had to make changes in the script.

In choosing words to prepare vocabulary questions, Tiger, Tahiri and Black Eagle tried to choose the words they emphasized a lot in their classes. They also made personal judgments about the words in terms of whether they were easy or difficult and whether they liked the word while they were trying to select. Tiger and Black Eagle also chose the words from the course book exercises randomly. In addition, Tiger tried to select the words which he found tricky.

4.3.2.1.1.3. Deciding what to ask

Tiger, Aziz, Beşiktaşlı, İlkbahar and Black Eagle had to decide what to ask from a listening audio and/or reading passage in their midterm exams. They first read and/or listened. During their listening and/or reading, they considered a piece of information important, so they prepared a question for that piece of information. They tried to find a piece of information with which they could assess a certain reading or listening skill. In addition, they tried to find a piece of information for which they believed they could prepare a question. The following excerpt of Aziz from the think-aloud protocol clearly illustrates such procedures which he used in writing his second midterm listening question.

Aziz (TAP): I am going on writing the second question. Now, there are pieces of general information about the life of Jason Stone, the place he died and how old he was in the part that I have listened. The ones which have caught my attention most among these pieces of information in the part I have listened are the place where he died, how old he was when he died and his constant business trips. They have caught my attention a lot. I have to empathize my students when I listen to something. When I listen, I have to determine which part my students can understand better and where the speaker emphasizes a piece of information. Therefore, my second question will be about where he died. My second question is "Where did he die?". It is in his London home. When I look at the previous answer, it is the option b, so I am thinking of writing the correct answer in the option b. When I say London home here, where could he die? Which place comes to my mind? It might be a place in the house. It might be the working room or office. I should mention them especially because I have asked my first question related to his job. I can use something related to his job as a distractor. Therefore, I will write in his office in the option a. I wrote the correct answer in the option b. I will write in his study which is completely unrelated in the option c. My aim is to check whether my students can listen for finding specific information to answer the question.

Tahiri also paid attention to the amount of time he spent on a topic in his class. If it was a lot, he used that topic in deciding what and how to prepare his questions. In addition, Tiger decided what to ask by taking into account his students' attitudes toward the parts of his lessons like the questions in the get ready parts of his course book. His students did not answer these questions in his classes, so he decided to ask such questions in his midterm exam. Though Crazy Soul used his previous midterm exam without changing, he spent some time on whether he should ask questions from his students' presentations, but he decided not to ask because he wanted his students to understand the importance of the presentation which they ignored.

4.3.2.1.1.4. Preparing the questions

In terms of preparing the exam questions, each participant preferred to use either writing their own questions or using the available ones, or both. In writing their own questions, Beşiktaşlı, Black Eagle, Tahiri, Tiger, İlkbahar and Aziz wrote their questions on their own and/or used their course books and/or benefitted from other sources. In using the available questions, İlkbahar, Tiger, Crazy Soul, Beşiktaşlı, Black Eagle and Deniz used the test book of their course books, the available questions in their course books without changing, the whole of their old exams or some parts of them, the ones on the websites without changing and/or the questions prepared by another colleague. How each participant prepared their own exam questions was explained below:

1. **Beşiktaşlı:** He wrote his own listening and reading questions. He used the vocabulary part of his previous midterm exam.
2. **Tahiri:** He wrote his own grammar and vocabulary questions. In addition, he used a dictionary and grammar book to write some of his grammar and vocabulary questions.
3. **Deniz:** She selected her midterm exam questions from the test book of her course book. She only added the fourth options to the questions.
4. **Black Eagle:** He prepared his own listening questions. In writing his vocabulary questions, he wrote some questions on his own, developed some based on the course book exercises and used some course book exercises as his exam questions without changing them.

5. **Tiger:** He mainly used his course book to write his vocabulary and open-ended questions. He used the exercises in his course book without changing them or with making small changes in them. In addition, he wrote some vocabulary questions on his own. He prepared his listening questions by himself.
6. **Crazy Soul:** He used his previous midterm exam without changing it.
7. **Aziz:** He developed his listening and reading questions on his own.
8. **İlkbahar:** She used the available listening questions on the website and also wrote her own questions. She evaluated the reading questions prepared by her colleague and used them with making some changes. She used the vocabulary questions of her previous midterm exam.

During writing their questions, each participant continuously brainstormed and outlined. They tried to decide the number of questions, the types of questions, the content of the questions, timing for the questions and the weights of different skills or sections in their exams while brainstorming and outlining. The quotation of Black Eagle from the think-aloud protocol clearly exemplifies these procedures.

Black Eagle (TAP): How many questions can I ask from this dialogue? I will check how many questions I can ask from this. Actually, there is not a limitation on the number of the questions. There may be three questions or five questions. However, I will try to ask as many questions as possible from the dialogue. OK! This is the topic. There is a product. It is a defect one and causes a problem. What type of questions can I use here? I may use multiple-choice and true-false questions. Let's start with two true-false questions.

In addition, Aziz and Beşiktaşlı brainstormed and outlined the number of the options in their exams. Both of them preferred using three options with listening questions and four options with reading questions because they thought that their students were not good at listening, so they wanted to make their listening questions easier than their reading questions.

Tiger, Aziz, Beşiktaşlı, İlkbahar, Black Eagle and Tahiri wrote all, most, or some of their questions in their exams on their own. They first talked to themselves about what to ask, second brainstormed about the content of the exams in their private speech and third code-switched in writing their own questions. They code-switched in

writing either the stem or the options, or both. The following quotations help to understand the procedures used by the participants in writing their own questions.

Beşiktaşlı (TAP): Our first question is generally what the passage is about. What is the passage mostly about? The passage generally mentions neuro-marketing. It is finding out the clients' brand choices by obtaining their reactions in their brain when they see brands related to a type of product through placing electrodes on their heads. I am writing about this. Its correct answer is, A, a new method to learn consumer choices. Generally, a new way or method of learning consumers' choices.

Tahiri (TAP): The second one is to encourage. What can I write for it? I mentioned the structure 'encourage someone to do something' and made them [his students] write their own sentences. Therefore, I should definitely ask it. Let's do it like this. Their departments are related to teaching. Therefore, I should write a sentence related to being an effective teacher. An effective teacher should --- their students to participate to... Is 'participate' used with to or in? Participate to or participate in? Yes, an effective teacher... They can learn a feature of an effective teacher. An effective teacher encourages his students to participate in classroom activities actively. This is a good one.

In addition to these procedures, the participants used some other procedures. Firstly, Tiger, Aziz, İlkbahar and Beşiktaşlı wrote their instructions for the parts which they would start preparing questions for, but Black Eagle and Tahiri wrote their instructions after they were done with writing their questions. Secondly, Tiger, Aziz, Beşiktaşlı, İlkbahar, Black Eagle and Tahiri often referred back to what they had taught in their classes while writing their questions. They also used the types of the questions (like matching, fill-in-the-blank, true-false and multiple-choice) which were similar to their classroom activities and with which their students were familiar. They took into account whether their students could understand and answer their questions while preparing their questions. Similarly, İlkbahar paid attention to her students' comments about her previous exams, so she tried not to prepare her questions in the way that her students complained about. Tiger cared his students' motivation when he wrote his questions. Like İlkbahar, Black Eagle tried not to ask any question about which his students might complain after the exam. Despite this, they also made their exams challenging enough for their students by adding an extra word or option because they wanted to see who studied and did not study at the end of their exams. Tiger, Aziz, Beşiktaşlı, İlkbahar, Black Eagle and Tahiri related the content of their grammar, vocabulary, listening and/or reading questions to the topics which they had taught in

their classes upon writing the questions. In addition, Tahiri related his questions to his students' daily lives and future professions.

In writing the options, Beşiktaşlı focused on the lengths of his options. In addition, he and Aziz thought that there would be two options which were very close to each other in terms of the correct answer, one option which was not related to the correct answer and one option which was not related to the correct answer, but close to it. While Aziz distributed the correct answers among the options like A, A, B, B and so on, the others placed them among the options randomly.

When preparing the midterm listening and/or reading questions, Tiger, Aziz, Black Eagle, Beşiktaşlı and İlkbahar ordered them according to the order of the events in the audio and/or reading passage. Beşiktaşlı and Aziz paraphrased the options in their reading questions. Tiger, Tahiri and Black Eagle ordered the vocabulary questions randomly. As Tahiri and Black Eagle wrote multiple-choice vocabulary questions, they paid attention to using the same parts of speech in their questions and options.

Tiger, Black Eagle and Tahiri prepared their own vocabulary questions. They wrote their own sentences in some questions. In addition, they used other ways to write their vocabulary questions. Tahiri preferred using a dictionary to write the definitions for his matching questions and to write multiple-choice fill-in-the-blanks questions. Black Eagle used his classroom examples and the ones in his course book to prepare his multiple-choice vocabulary questions' stems. He and Tiger chose the words and their definitions from the matching exercises in their course books and used them as their midterm matching questions without making any change. Tiger also looked at other vocabulary exercises in his course book, chose some and used them as his midterm vocabulary questions either without making a change or with making small changes. He also took the definition of a word in a matching exercise from his course book and converted it into a fill-in-the-blank question.

4.3.2.1.1.5. Self-assessing the written questions

After preparing one of or all of their questions, Aziz, Beşiktaşlı, İlkbahar, Black Eagle, Tahiri and Tiger self-assessed the questions. In their self-assessment, they checked whether the stems could be understood by their students, whether the wording

and the use of grammar were correct in the stems and the options, whether the questions could assess what they wanted it/them to assess and whether the answers of the questions were prepared correctly. Black Eagle's excerpt below clearly explains these procedures.

Black Eagle (TAP): For example, I can ask it. It can be similar to the example that I gave in the class. For instance, a famous singer may sue against a newspaper. Why? Because of law... For example, I can say "Every day we read in the newspapers that one of the celebrities, celebrities, every day we read in the newspapers that one of the celebrities sue against, one of the celebrities sues against a." We read in the newspaper that he/she sued against a newspaper. This sounds a little weird. Or we can say we hear. Every day we hear that. We hear that one of the celebrities sues against a newspaper or magazine, a magazine or newspaper, because of ... What is the correct answer? What should I say in the option a? Or where should I write the correct answer? We have four options: a, b, c and d. For example, I should write the correct answer in the option b. Deformation, blackening someone. He/she sued because of deformation. The answer is deformation. We chose and asked a noun. We should use nouns in the other options. For example, we use intent meaning willingness. What else can I use? What else can I use? Let's look at the other units. We can use notion. I used notion because it is a noun. Another noun? I can use movement. Let's check other options whether they can also be answers. Movement... Because of the person who wanted, he/she sued. He/she sued because of deformation. The right answer is b. It cannot be the answer that he/she sued because of intention and notion.

In addition to these procedures, the participants self-assessed the similarity of their exam questions' types to the ones used in their classes. If they were sure that their questions met these criteria, they decided to use them in their exams. In this way, Tahiri evaluated the quality of his questions and if he was satisfied with the result of his self-assessment, he reinforced himself.

4.3.2.1.1.6. Evaluating the available questions

Some participants preferred using the available questions. Beşiktaşlı, İlkbahar and Crazy Soul decided to use their last year midterm vocabulary exam questions in their new midterm. They first checked whether they used the same syllabi and course books in the previous term. Then, they matched the units for which the old vocabulary questions were prepared with the ones for which the new vocabulary questions would be prepared. As a result of this, they decided to use the questions covered by the units from which both the old and new vocabulary questions were chosen and they omitted the others. They also thought whether they had experienced any problem when they had

used the questions in their previous exams, which was also an effective criterion for their decisions. As the previous midterm exams' vocabulary questions met these criteria, they decided to use them.

In addition, İlkbahar decided to use the available listening questions on the website and the reading questions which one of her colleagues had given to her. Her exam included both listening and reading. She checked whether the questions were understandable and answerable to her students. Once she understood that they could understand and answer the questions, she decided to use them in her exams.

Like Beşiktaşlı and İlkbahar, Crazy Soul decided to use the available listening and reading questions. Those questions were the listening and reading questions of his previous midterm exam. He self-assessed the listening audio and reading passage in terms of the criteria through which some participants chose their listening audio and reading passages to prepare their listening and reading questions was explained. Then, he checked whether the listening and reading questions were similar to his classroom activities and whether his students could understand and answer them. He decided that they were suitable to use in his new midterm exam, so he used them without making any change. He did so because of the lack of time, the levels of his students and using the same syllabus and course book.

Similarly, Deniz chose her midterm questions from the test book of her course book because she said that she was not an expert on preparing questions, it was time-saving, it provided content-validity and she did not experience any problem with the course book, its exercises and its answer keys before. She also paid attention to her students' levels of English, her testing environment and the testing program, Blackboard she used to prepare and grade her exams. She chose the questions related to what she had taught in her classes. She chose her listening audio and reading passage whose topics were similar to the ones used in her classes. In the beginning, she determined the number of the questions for each skill. In addition, she added one more option to the questions in order to make them more challenging.

4.3.2.1.1.7. Finalizing the preparation of the exams

In the final stage, each participant self-assessed their questions again. The ones who prepared listening and/or reading questions also self-assessed the listening audio and/or reading passages they chose. In self-assessing the listening audio, the participants paid attention to their durations, recordings, topics and understandability. Beşiktaşlı's quotation below is an example of how these participants self-assessed the listening audio they chose and the types of questions they used in their midterm exams.

Beşiktaşlı (TAP): The things that we [teachers] pay attention to in preparing listening questions is finding a related listening audio. It is suitable for their [students'] levels, easy for them to understand, related to classroom topics and includes a lot of words related to the classroom topics. Besides, the questions I prepared from the listening audio are the question types that my students have practiced in my classes and in the course book and that they are familiar with.

In assessing their reading passages, they paid attention to whether the passages included the words taught in the class, whether they were understandable for their students and whether they were similar to the classroom reading passages in terms of their topics and lengths. In addition, Aziz assessed his instructions and the question words he used in his questions to be sure that his students could understand them. As a result of such self-assessment, Tahiri made some changes in his questions. Similarly, İlkbahar and Tiger self-assessed the variety of the questions in their exams.

In addition, some participants preferred preparing their answer keys while writing their questions while others prepared their answer keys after writing their questions. The participants checked the number of the questions they prepared for each skill and wanted to make sure that the numbers reflected the weights given to the different skills in their classes. They also scored their questions in the end. Tahiri, Aziz, Black Eagle, Beşiktaşlı, Deniz and Crazy Soul scored their exams depending on the number of their questions. İlkbahar gave scores to her questions according to the weights of EAP and ESP in her classes, while Tiger scored his questions based on their difficulty levels.

Finally, Beşiktaşlı, Tahiri, İlkbahar, Aziz, Tiger and Crazy Soul typed their exams on their computers. On the other hand, Black Eagle preferred writing his questions on a paper and Deniz chose her questions from the test book of her course book. They put them in their drawers and locked their drawers.

4.3.2.1.2. The participants in documents

The participants' quizzes, midterm and final exams were content-analyzed depending on the codes indicated in Table 4.7 below.

Table 4. 7.

Codes of Document Analysis

-
1. Practicality
 2. Reliability
 3. Content validity
 4. How the questions were formed
 - 4.1. Reading passages and questions
 - 4.2. Listening audio and questions
 - 4.3. Vocabulary questions
 - 4.4. Grammar questions
 - 4.5. Open-ended questions
 5. Face validity and bias
 6. Authenticity
 7. Washback
-

As this part focuses on how the participants prepared their exams only, the analysis related to the forming questions was detailed. The other three codes were previously explained in this chapter.

4.3.2.1.2.1. Reading passages and questions

Crazy Soul, Beşiktaşlı, Aziz and İlkbahar totally used two reading passages in their midterm and final exams, while Tiger used a reading passage in his midterm exam. The passages they chose for their exams were similar to the ones which their students had studied in their classes, but Crazy Soul used a passage whose topic was not similar to the ones in his course book in his final exam. They used true-false and multiple-choice questions which were similar to the classroom activities, but Tiger preferred using two open-ended questions in his midterm exam. In addition, their questions assessed the reading skills like scanning, scimming, finding the main idea and comprehending which were practiced in the reading sections of their course books, but Crazy Soul used a summarization question which was not practiced in his course books in his midterm exam.

4.3.2.1.2.2. Listening audio and questions

Crazy Soul, Tiger, Aziz, İlkbahar and Beşiktaşlı chose and used two different audio in their midterm and final exams, while Black Eagle prepared his own scripts for his midterm and final exams. In addition, Tiger used one of the listening audio which his students studied in his classes in his final exam. They except for Crazy Soul used the audio whose topics were similar to the ones of the listening audio in their course books because Crazy Soul used an audio which was not similar to the classroom listening audio in terms of its topic in his final exam. Aziz prepared only multiple-choice listening questions in his midterm exam and true-false listening questions in his final exam. Tiger used an open-ended fill-in-the-blank listening part in his midterm exam and used the questions of the audio he had chosen from his course book in his final exam without changing the questions. Their questions assessed the listening skills like listening for finding specific and general information which were practiced in their classes. The participants used the types of the questions which were similar to the classroom activities in their exams.

4.3.2.1.2.3. Vocabulary questions

Beşiktaşlı, Tahiri, Black Eagle, Tiger, Crazy Soul, Aziz and İlkbahar prepared vocabulary questions in their exams. They chose their words from their course books. Tiger asked open-ended fill-in-the-blanks questions, while the others asked such questions as multiple-choice questions. In preparing such questions, Beşiktaşlı, Tiger and Crazy Soul chose their questions from the course book exercises and used them either without changing or with small changes. In addition, Beşiktaşlı, Tiger, Crazy Soul and İlkbahar took some definitions from the matching exercises of their course books and converted those definitions into multiple-choice questions. Beşiktaşlı, Tahiri, Tiger, Black Eagle and İlkbahar also wrote their own sentences for such questions. They also used matching in the vocabulary sections of their exams. In preparing their matching questions, İlkbahar, Black Eagle, Beşiktaşlı and Tiger chose their words and definitions from the matching exercises in their course books and used them without changing. In addition, Beşiktaşlı and İlkbahar took the definitions of some words from the glossary parts of their course books and used them without changing. Tahiri used

the definitions given in a dictionary, while İlkbahar preferred writing her own definitions for some words. In terms of matching questions, İlkbahar used four different types of questions: (a) finding the words whose definitions were given among a group of words, (b) matching the words with their definitions, (c) deciding whether the words and their definitions given were true or false and (d) choosing the words whose definitions were given in the stems among four options. In addition, she and Tahiri prepared a cloze test based on the topics of their lessons in their midterm exams. Meanwhile, Tahiri used some midterm vocabulary questions in his quiz. Crazy Soul prepared different vocabulary questions in the vocabulary parts of his midterm and final exams, but some of his vocabulary questions had the same options used for different stems. The types of the questions the participants used in their exams were similar to their classroom activities.

4.3.2.1.2.4. Grammar questions

Tahiri prepared grammar questions in his quiz, midterm and final, while Beşiktaşlı and İlkbahar prepared grammar questions in his midterm and her quiz. Tahiri generally used multiple-choice and matching in the grammar sections of his exams. He reused some of his midterm grammar questions in his quiz. He and Beşiktaşlı used cloze test once in their exams. İlkbahar preferred fill-in-the-blanks as constructed response in her quiz. Their students were familiar with them because of their classroom activities. The questions assessed only the grammar topics taught in their classes.

4.3.2.1.2.5. Open-ended questions

Deniz, Tiger, Black Eagle and Beşiktaşlı used open-ended questions in some of their exams. Deniz prepared her open-ended questions based on the titles of the sections in her course pack, so she related them to her course objectives. Tiger preferred using the available questions in the get-ready part of his course book separately or by combining them with each other. Black Eagle and Beşiktaşlı developed their open-ended questions based on the main idea or the content of the units in their course books.

4.3.2.2. The types of the questions used in the chosen assessment methods

In terms of the types of the questions in the selected response exams, all participants preferred to use multiple-choice, true-false, matching, cloze test and fill-in-the-blanks because these types of questions were practiced in their classes and their students were familiar with them as understood from the second interviews. In addition, Beşiktaşlı thought that such types could decrease his students' anxiety and fear because of their similarity to his classroom activities and of his students' familiarity with them. Besides, İlkbahar wanted to vary the types of the questions in her exams in order to see what her students learned and did not learn. On the other hand, Aziz chose them because he wanted to reduce his workload due to the number of his students in his classes and his teaching hours. Deniz, Tiger and Crazy Soul used different types of assessment methods for formal assessment. Deniz used essay writing and presentations as well as the types mentioned at the beginning of this paragraph in her courses because she thought the names of her courses and their content required her to use them. Like her, Crazy Soul used presentation as performance assessment in his classes because he wanted to help his students improve their English, practice it and understand that it is necessary for their majors. Tiger used fill-in-the-blank and open-ended questions as well as other types of questions like multiple-choice and matching in his exams in order to see whether his students could use what they learned. On the other hand, Tahiri preferred using sentence writing as a type of constructed response in order to help and motivate his students to learn.

4.3.2.3. Providing validity

In terms of making the exams valid, Beşiktaşlı, Tahiri, Deniz, Black Eagle, Crazy Soul and İlkbahar checked what they studied in their classes and prepared their questions accordingly as the second interviews revealed. In accordance with this, Crazy Soul tried to choose his exam reading passages and listening audio which were similar to the classroom ones in terms of their lengths, topics, appropriateness to the level of his students, the number of the words and including the target words. İlkbahar told that she used different types of questions and followed her assessment criteria strictly in order to make her exams valid. Beşiktaşlı, Tiger and Aziz used peer interaction to make their

exams valid. They told they gave their exams to their colleagues, wanted them to check their questions, received feedback from them and made changes if necessary. Deniz solved her questions again in order to determine whether what she could understand from them would be similar to what her students could understand by putting herself in their shoes while solving the questions. Similarly, Tiger piloted the types of the questions that he wanted to use in his midterm and final exams in his quiz. He also checked whether his students could understand his questions before his exams. Deniz also used essay writing and presentation in her classes. To make her essay exams valid, she told she chose two topics for which her students could produce ideas. To make her presentation assessment valid, she told she made her students prepare an outline, check their outlines with her and decide what should be and should not be in the presentation. Like Deniz, Crazy Soul also used presentation in his classes. He taught his students how to prepare an effective presentation, gave them a list of topics related to their majors to choose and used a rubric to grade his students in order to make his presentation assessment valid.

The focus group discussion indicated that each participant paid enough attention to the content validity of their exams in preparing their questions. For example, Black Eagle told that he opened his course books, looked at the units and decided from which units he would prepare questions before writing them. In terms of how to prepare questions, Crazy Soul, Deniz and Tiger said that they preferred using the available questions on the Internet, their test booklets and/or their course books with making some changes because they did not have enough time to prepare new and original questions. Beşiktaşlı added that he and his colleagues tried to use the types of the questions similar to the ones used in their course books in their exams. By doing so, Tiger believed he provided the consistency between his exams and course book. In addition, Deniz emphasized preparing understandable and clear questions for her students to answer if she prepared her own questions. Crazy Soul also said he paid attention to the lengths of the reading passages and listening audio in his exams and tried to use the ones similar to the ones used in his classes in terms of these aspects. He also added that lack of time, the number of their students and the number of the classes the participants had to teach prevented them from writing original questions.

The document analysis demonstrated that each participant clearly defined the units which each exam would cover. Their questions were related to what was studied in their classes. They used the types of the questions which their students were familiar with. They related their questions to their course objectives. Deniz prepared a guideline for her students to prepare their presentations in accordance with her course content and classroom activities. Beşiktaşlı, Tiger, Aziz and İlkbahar chose and used the listening audio and/or reading passages whose topics were similar to their classroom reading and listening topics. Crazy Soul chose and used an audio and passage which were similar to the ones used in his classes in his midterm exam, but not in his final exam. Black Eagle wrote his own scripts by integrating two or three topics in his course book. Tahiri, Beşiktaşlı, Black Eagle, Tiger, Crazy Soul and İlkbahar chose the target words which their students learned in their classes to prepare their vocabulary questions. In addition, Tahiri, İlkbahar and Beşiktaşlı prepared their grammar questions based on the grammar topics which they had covered in their classes in all or one of their exams. The participants did these things to make their exams content valid.

In addition, İlkbahar, Aziz, Tiger, Black Eagle and Deniz used clear, short and understandable instructions in their exams. However, Crazy Soul did not use any instruction in his midterm and final exams. Beşiktaşlı and Tahiri used such instructions in one or two of their exams, but did not use any instruction in the rest of their exams. The participants prepared their questions according to their students' level of English by using the types of the questions which their students were familiar with and the language in their questions similar to the one used in their course books. Beşiktaşlı, Black Eagle, Tiger, Aziz and İlkbahar used the listening audio and reading passages whose topics were similar to the ones their students had studied in their classes in their exams, but Crazy Soul did not use such an audio and passage in his final exam though he used such things in his midterm exam. Besides, the lengths of the audio and reading passages were similar to the ones of the course book reading passages and listening audio. Finally, Tiger, Aziz, Crazy Soul and Black Eagle followed a logical organization and structure in their exams, while Tahiri did not follow. In addition, Deniz, Beşiktaşlı and İlkbahar followed a logical organization and structure in one or two of their exams, but not in the rest of their exams or in some parts of their exams. The participants used these strategies for face validity.

4.3.3. Administering exams, scoring them and interpreting their results

This part gives information about how the participants administered their exams, scored them and interpreted their results depending on the codes in Table 4.7.

Table 4.7.

Codes of the Third Sub-component of LAL

3. Administering exams, scoring them and interpreting their results
3.1. Definition of reliability
3.2. Providing reliability
3.3. Administering exams
3.4. The problems encountered in administering the exams and the ways used to overcome the problems
3.5. Scoring
3.6. Grading
3.7. Consistency of assessment interpretation
3.8. Interpreting formal and informal student evaluation
3.9. Using student assessment results for assessment tools and students' learning
3.10. Washback effect
3.11. Confidentiality of assessment and assessment results
3.12. Attitude towards exam complaint

In addition to the second interviews, the third and fourth observations were also analyzed in order to explain how the participants administered and graded their exams. The third observations were analyzed according to the codes in Table 4.8 below.

Table 4.8.

Codes of the Third Observations

1. Timing
2. Informing
3. Avoiding cheating
4. Not being distractive

Besides, the fourth observations were analyzed depending on the codes in Table 4.9.

Table 4.9.

Codes of the Fourth Observations

-
1. Double-checking
 2. Development of the grading tool
 3. Grading
 4. Online personal/public announcement
 5. Confidentiality
-

As the third and fourth observations were used to support the findings of the second interviews, their analyses were integrated and presented with the analyses of the related parts in the second interviews.

4.3.3.1. Definition of reliability

Tiger thought that reliability is “that questions are relevant to the subjects.” Similarly, İlkbahar defined reliability as “preparing an exam related to what students have learned.” She also added “preparing the exam suitable to their [students] levels and being valid in accordance with the variety of questions and quality” to her definition of reliability. Aziz also emphasized that reliability is “an exam that I can say will be valid, dependable and successful.”

While İlkbahar, Tiger and Aziz related reliability to content validity, Black Eagle and Deniz emphasized the quality of the exams prepared. According to Black Eagle, reliability is related to “doing an evaluation without pressure and any other external factor like a mistake in the answer key.” Like him, Deniz stressed out that taking any precaution against the things that may affect an exam negatively and double checking the exam are reliability as understood from the excerpt below.

Deniz (I2): I double check the assessment tool when I prepare it. I take precautions depending on who is assessed, physical conditions and the need for an extra instructor in relation to the number of the students in order to increase the reliability of my exam. I double check myself in terms of knowledge. I believe that all of these make my exam reliable.

In addition to the different definitions of Black Eagle and Deniz, Tahiri made a different definition of the term. He defined reliability as “giving equal points to each question or different points to questions depending on their difficulty levels.” Likewise, Beşiktaşlı and Crazy Soul made different definitions of reliability. According to Crazy

Soul, it is having the expected results from his students. That is, reliability is that the result he takes from an exam reflects his students' real performances as the quotation below indicates.

Crazy Soul (I2): Reliability is that an assessment tool gives reliable results. Does it [the assessment tool] give the results that we expect from our students? Are the results we take reliable? If a student gets 70 from the exam, does this grade show his real performance? If the assessment tool assesses what we [teachers] expect, then we can say it is reliable. Do the results it provides represent the quality of the students? This is what reliability is in my opinion.

As Crazy Soul defined reliability, Beşiktaşlı told reliability is “the suitability of our [teachers'] results to our goals that we have after exams and having the same results under the same conditions.”

Though some participants made similar definitions of reliability, they understood reliability differently. Their definitions focused on content validity, the quality of the exams prepared, types of the questions, meeting the expectations from students and suitability to the goals.

4.3.3.2. Providing reliability

For the exam reliability, Black Eagle, Tiger, Aziz and İlkbahar checked whether their exams covered what they taught in their classes as the second interviews showed. In addition, İlkbahar considered whether her questions were related to her course goals. Besides content validity, Aziz also took into account the face validity of his exams, that is, whether his students could read and understand his questions. Secondly, Crazy Soul, İlkbahar and Aziz benefitted from peer feedback in making their exams reliable. They shared their exams with their colleagues and wanted them to check the questions in terms of the use of language and their answers. Depending on peer feedback, they made changes in their questions if necessary. Thirdly, Crazy Soul and Deniz mentioned that they took necessary precautions to avoid their students' cheating in the exams. Fourthly, Deniz, Crazy Soul and İlkbahar double checked their exams after they prepared them. Fifthly, Crazy Soul and Black Eagle paid attention to the correct and objective grading of their exams. Besides, Beşiktaşlı matched his students' exam scores with their real performances which he formed depending on his in-class observations. Though Tahiri believed that the questions should be scored according to their difficulty levels in order

to make his exams reliable, he could not do so because of the mutual decision on how to score their exams (scoring according to the number of the questions).

As the think-aloud protocol findings revealed, each participant made their exams content valid to make their exams reliable. They also used the types of the questions which their students were familiar with. Besides, Deniz, Black Eagle, Tiger and Crazy Soul determined the weights of the different sections like listening and grammar in their exams depending on the importance given to each section in their course books; Tahiri and Aziz did it depending on the pre-determined criteria (asking 10 questions for each section); Beşiktaşlı and İlkbahar did it according to the distribution of ESP and EAP in their three-hour academic English course. Moreover, the participants used understandable and clear audio and/or reading passages whose topics were similar to the ones studied in their classes to make their exams reliable.

As found in the document analysis, the participants gave enough importance to the reliability of their exams. For this purpose, they first clearly printed their exams, so the questions in the exams were readable, but there was a problem with true-false options in Aziz's quiz. They informed their students about what they should study for their exams and how their learning would be assessed. In addition, Beşiktaşlı, Black Eagle, Tiger, Aziz and İlkbahar used clear and understandable reading passages and/or listening audio whose topics were similar to the classroom ones in their exams. Like them, Crazy Soul used clear and understandable audio and passages in his exams, but his final listening audio and reading passage were different from the classroom ones in terms of their topics. Moreover, the participants had objective scoring systems for their exams. When open-ended questions were used, Deniz, Beşiktaşlı, Black Eagle, Tiger and İlkbahar prepared detailed answer keys for such questions in order to grade them objectively. In addition, Deniz assessed her students' presentations twice at home and school. While Deniz, Tiger, Aziz and İlkbahar used clear, short and understandable instructions similar to the classroom ones in their exams, Beşiktaşlı and Tahiri used such instructions in some of their exams. Crazy Soul did not use any instruction in any of his exams. The participants prepared their questions suitable for their students' levels of English by using clear and understandable audio and/or reading passages similar to the classroom ones in terms of their topics (though one of them did not so in his final exam), by preparing the types of the questions which their students were familiar with,

by writing clear, short and understandable instructions (though some participants did not pay attention to this) and by using language similar to the one used in their course books. They also gave their students enough time to study their exams and their timing for their exams was appropriate. Besides, Beşiktaşlı, Black Eagle, Tiger, Crazy Soul and Aziz followed a logical organization and structure in their exams, yet Deniz and İlkbahar followed such an organization and structure only in some of their exams. Unlike them, Tahiri did not follow such an organization and structure in his exams.

All participants were observed to take some precautions in order to avoid cheating while administering their exams in the third observations. They thought that such precautions supported the reliability of their exams. First, they had their students turn off their cell phones and remove their notes from their desks. Second, they made them sit in two rows by leaving one row empty between them. Third, they informed them about the exam rules, how the exams would start, when the exams would start and finish and how they would shade their answers on their bubble sheets. Fourth, Beşiktaşlı, Deniz, Crazy Soul, Tiger and Aziz walked around the class during the exam silently; İlkbahar stood in front of the class where she could see her students; Black Eagle shared the responsibility with the other proctor and monitored the students sitting in front of the class. Fifth, the participants paid attention not to disturb the students during the exams.

The findings from the fourth observations showed that all participants double-checked their answer keys by answering their questions again and comparing their new answer keys with the old ones in order to make their exams reliable. They also compared the grades on the excel sheets or students' exam papers with the ones on the online information system before announcing them. In addition, they calculated their students' total grades twice. Beşiktaşlı, Tahiri, Black Eagle, Crazy Soul, Aziz and İlkbahar double-checked the answer keys on their grading tools with the ones they drew on the students' bubble sheets. They counted the number of their students' correct answers twice. In addition, Deniz, Tiger and Beşiktaşlı did not look at their students' names in order not to be affected while grading their papers.

4.3.3.3. Administering exams

The second interviews pointed out that Deniz and Crazy Soul prepared two types of booklets and asked for another proctor before starting to administer their exams depending on the number of the students in their classes. Beşiktaşlı also wanted one of his colleagues to proctor in his exams with him. Crazy Soul had to divide his classes into different sections because of the number of his students, so he announced the class lists one week before the exam date. When starting to administer their exams, Beşiktaşlı, Black Eagle, Tiger, Aziz, İlkbahar and Tahiri first informed their students about the exam, its rules, its duration and how to answer the questions. Before this, Deniz and Aziz read their students' names on their lists to be sure that they were in the right class to take their exams. Black Eagle, Beşiktaşlı, Deniz, Crazy Soul, Aziz and İlkbahar secondly arranged their students' seating orders. Meanwhile, Beşiktaşlı, Deniz and Aziz wanted their students to turn off their mobile phones. Beşiktaşlı also checked his students' desks to see whether there was any document. Then, Black Eagle, Tiger, Aziz and İlkbahar distributed their exam papers to their students and wanted them to check whether there was any problem. In addition, if their students had any question about the exam to ask, they answered their questions. İlkbahar wrote the exam duration on the board. The participants paid attention to starting their exams on time. During the exam, they walked in the classroom silently to help their students if they had any question about the exam and to avoid cheating. Black Eagle, Tiger, Deniz and Crazy Soul tried not to disturb their students during the exam. Tahiri tried not to stand in front of a student a lot in order not to disturb him. Meanwhile, Deniz wanted the other proctor to take attendance while İlkbahar took the attendance on her own. Beşiktaşlı, Crazy Soul, Black Eagle and Tiger reminded their students of the remaining time. Tahiri warned his students about signing the attendance list. If some students finished their exams earlier, Tiger, Deniz, Aziz and Crazy Soul wanted them to put their exams on their tables and leave the classes without making any noise. Tiger, Crazy Soul, Aziz and İlkbahar finished their exams on time. Tahiri, Beşiktaşlı, Black Eagle, Crazy Soul and Tiger collected their students' papers when their exams were over.

Each participant told in the focus group discussion that they followed a standardized procedure in administering their exams. They said they informed their

students about the types of the exams, exam rules and what to do in different sections. They seriously explained the exam rules in order to avoid any cheating issue. Then, they distributed their papers and wanted their students to check the papers if there was any problem and they answered their students' questions once the students asked any. They also reminded their students of the remaining time. They generally made their exams in big classes with two or three instructors due to the excessive number of the students in their classes. They aimed to make their assessment and evaluation sound. These things are related to administering paper-based exams. Deniz and İlkbahar also administered computer-based exams in some of their classes. They followed a similar procedure to administer computer-based exams. They made their students sit according to their levels of English as elementary and pre-intermediate by leaving a chair empty between them. They explained to their students how to answer the listening and reading questions on their computers. They had two or three more proctors to help their students because the students had a lot of questions to ask during the computer-based exams. They claimed that this type of administering an exam was busier than administering a paper-based exam.

The third observations showed that the participants went to their exam venues on time except Tiger. Meanwhile, Crazy Soul stuck the student lists on the exam venues' walls and told the students to find their names on the lists as the number of the students who attended his exams were a lot. The participants made their students sit in two rows by leaving one empty row between two rows, turn off their cell phones and remove their notes on their desks. They informed their students about the exam questions, how their exams would, when they would start and finish and how the students should shade their answers on their bubble sheets. They distributed their exam papers to their students and answered their students' questions if the students had some questions. They wanted their students to write their names and student ID numbers on the exam papers and Crazy Soul also wanted them to write their course teachers' names because he, Black Eagle and Aziz thought academic English courses to the faculty of architecture students. Black Eagle, Crazy Soul and İlkbahar prepared two types of booklets (A and B) and distributed their papers as A-B-A-B for one row and B-A-B-A for the other row. Beşiktaşlı wanted his students to check his exam papers in case of a printing mistake. Meanwhile, Tahiri, Black Eagle and Crazy Soul's exams started with listening, so they

wanted their students to read the listening questions, to give feedback about whether they could hear the audio and to follow their instructions like the first listening starts and finishes. They checked the audibility of their audio and played it twice. In the meantime, some students came to Tahiri, Deniz and Crazy Soul's exams late, but they allowed the latecomers to take their exams. They explained the rules to such students silently. In addition, the participants wanted the students to sign the attendance lists and warned the ones who did not sign after comparing the number of the signatures with the number of the students in the exam venues. During the exam, Beşiktaşlı, Deniz, Tiger, Crazy Soul and Aziz preferred walking around the class; İlkbahar and Beşiktaşlı stood in front of the class; Tahiri walked around the class for some time, but he sat on the teacher table and spent his time dealing with his cell phone. While walking around the class, Tiger, İlkbahar, Deniz, Black Eagle, Tahiri and Beşiktaşlı helped some students who asked a question silently. Meanwhile, Crazy Soul was informed about a problem related to the exam. He stopped the exam, informed the students about the problem, had them make necessary changes and restarted the exam. Moreover, Tahiri, Deniz, Black Eagle, Tiger, Aziz and İlkbahar gave information about the remaining time to their students. The participants made the students who finished their exams earlier, but sat in the classrooms leave the exams by putting their papers on their desks and without making any noise. Tahiri, Beşiktaşlı, Black Eagle, İlkbahar and Aziz finished their exams on time without giving extra time to their students. However, Deniz could not do so because some students had clashes in their exam schedules and informed her about this situation before the exams, so she allowed such students to take her exams and gave them extra time. Tiger came to his exam venue late, so he had to start his exam a few minutes later than the scheduled time and to add this missing time to the exam duration to compensate. Crazy Soul also waited for one of his students to finish shading his answers on the bubble sheet.

4.3.3.4. The problems encountered in administering the exams and the ways used to overcome the problems

The participants, except Tahiri, told in the second interviews that they encountered cheating issues in administering their exams. In addition, Aziz said that some students wanted to go to the restrooms in the middle of the exams or some who

finished their exams earlier waited and talked in front of the class, while the rest were taking their exams.

As the second interviews indicated, Beşiktaşlı and İlkbahar tried to catch the cheating student's attention by standing near his desk to avoid cheating problems in administering their exams. Beşiktaşlı, İlkbahar, Aziz, Tiger, Crazy Soul and Tahiri warned him silently. Black Eagle, Crazy Soul, Aziz and İlkbahar changed his seats and took his cheating material. In addition, Crazy Soul and Aziz took his/her exam paper, while Black Eagle preferred talking with him/her after the exam. Besides them, Tahiri put a mark on that student's exam paper and cut some points. Tahiri, Crazy Soul and Tiger preferred writing a report about him to his department. In addition to these precautions, Aziz explained the exam rules to his students before the exam started and did not allow any of them to go to the restroom without giving their exam papers to him. All in all, Tiger and İlkbahar's precautions were shaped based on the types of cheating (looking at another student's paper and using a cheating material), while Crazy Soul and Aziz's measurements were determined according to the degree of the student's insistence on cheating.

The third observations revealed that the participants made their students remove their notes on their desks, turn off their cell phones before the exams started and sit in two rows by leaving one row empty between them. They also informed them about their exam rules in the beginning. Despite these precautions, Black Eagle, Tiger, Aziz, İlkbahar, Crazy Soul and Beşiktaşlı encountered cheating problems in their exam administrations. They went to the students who tried to cheat, warned them silently, changed their places and put some of the students' exam papers closer to them because the students sitting behind those students tried to look at their exam papers to cheat.

4.3.3.5. Scoring

During second interviews, İlkbahar, Beşiktaşlı, Tahiri, Black Eagle and Crazy Soul told that they scored their questions depending on the weights of different language skills (listening and reading) and parts of the language (vocabulary and/or grammar). Beşiktaşlı and İlkbahar designed their scoring systems based on the distributions of ESP and EAP (two hours for ESP and one hour for EAP) in their three-

hour academic English classes. İlkbahar gave three points to her ESP questions and two points to her EAP questions by asking the same number of questions in the EAP and ESP parts of her exams (20 questions for the EAP part and 20 questions for the ESP part). On the other hand, Beşiktaşlı gave 2.5 points to his EAP and ESP questions, but he asked 25 vocabulary questions in the ESP parts of his exams and 15 listening, reading and/or grammar questions in the EAP parts of his exams. Similarly, Crazy Soul reflected the weights of different language skills and parts of language in his exams by asking a different number of questions in different sections of his exams, yet he gave the same points to all questions as it made grading his exams easier. Like Beşiktaşlı and Crazy Soul, Black Eagle gave the same points to all questions, but he asked a different number of questions in different sections of his exams. Unlike them, he reflected the difficulty levels of different language skills and parts of language he determined on his own by doing so, but not by the weights given to each skill and part in his course book. However, Tahiri asked the same numbers of questions in his exams and gave the same points to them because of the mutual decision on the number of the questions to be asked in the midterm and final exams and of grading his exams easily. Deniz and Tiger scored their exams according to the types of the questions and their difficulty levels. Aziz scored his exams based on the types of questions as well as what he considered as important.

The first observation revealed that all participants agreed that they would ask 40 questions in their midterm and final exams. They would ask an equal number of questions for different sections if they taught only EAP, but if they taught ESP and EAP together, they would ask a different number of questions for ESP and EAP, but the total number of the questions had to be 40. This decision was made because different participants taught different things in their academic English courses. That is, Tiger and Black Eagle taught ESP; Aziz, Tahiri and Deniz taught EAP; Crazy Soul, İlkbahar and Beşiktaşlı taught ESP and EAP in their academic English courses.

In addition, the participants stated in the focus group discussion that they scored their questions according to the number of the questions in order to make grading their exam papers easy because giving different scores to different questions in different sections would be time-consuming and time was the thing they lacked. However, they believed that they should score their questions according to their difficulty levels.

İlkbahar added that the distribution of EAP and ESP in her three-hour academic English courses had an effect on scoring her ESP and EAP questions. Crazy Soul emphasized that they decided on the sections in their exams in accordance with the content of their course books and classroom activities. In addition, Black Eagle told that they focused on vocabulary more than other skills in their exams because their course books aimed to teach words more than other skills. Therefore, Crazy Soul said that they thought their classroom activities shaped their exams' structure. On the other hand, Black Eagle told he used open-ended questions in some of his exams and gave them high points because of their difficulty levels.

From the think-aloud protocol, it was found out İlkbahar and Beşiktaşlı determined the weights of the different sections in their exams depending on the distribution of ESP and EAP in their three-hour academic English classes (they studied ESP for two hours and EAP for one hour with their students). To show the distribution, Beşiktaşlı asked 25 ESP questions and 15 EAP questions by giving 2.5 points to each question, whereas İlkbahar asked 20 questions for ESP and 20 questions for EAP by giving 3 points to each ESP question and two points to each EAP questions. Tahiri and Aziz gave the same weights to the different sections in their exams owing to the pre-determined criteria (ask 10 questions for each section). They asked 10 questions for each section and gave 2.5 points to each question. In addition, Deniz, Crazy Soul and Black Eagle decided the weights of each section based on how much importance their course books gave to each section, so they asked different numbers of questions in each skill to reflect this distribution. Deniz asked 50 questions in total and gave two points to each question; Black Eagle asked 20 questions and gave each question 2.5 points (he and his partner prepared 40 questions in total and he was in charge of preparing 20 questions); and Crazy Soul asked 35 questions in total and gave 2.9 points to each question. Besides them, Tiger scored his questions according to their difficulty levels and gave the highest score to the vocabulary section as his course book gave more importance to vocabulary than the other sections.

In addition to the think-aloud protocol, the document analysis revealed that Beşiktaşlı, Deniz, Black Eagle, Tiger, Crazy Soul and İlkbahar prepared a different number of the questions in each section of their exams like listening and reading in accordance with the importance given to each section in their course books. Tahiri

prepared the equal numbers of the questions for two sections in his parts in the midterm and final exams, but his partner, Aziz prepared more listening questions than reading questions in his part in the final exam, while he prepared an equal number of the questions for the reading and listening parts in the midterm. In addition, Tahiri prepared more grammar questions than vocabulary questions in his quiz. Deniz and Tiger preferred scoring their exams according to the difficulty levels of the questions and to the importance given to the different sections in their course books. Deniz also developed her rubric for evaluating her students' presentations in relation to her classroom activities and the do's and don'ts of making presentations. On the other hand, Black Eagle and İlkbahar did not share any information about scoring in their quizzes.

4.3.3.6. Grading

As the findings of the second interviews demonstrated, Beşiktaşlı, Tahiri, Deniz, Black Eagle, Crazy Soul, Aziz and İlkbahar prepared their own grading tools by putting a transparent paper on a bubble sheet where they shaded their answer keys and drew the key answer keys on the transparent paper. Second, they put their grading tools on their students' bubble sheets, counted the numbers of the correct answers and multiplied them with the points given to the questions. During the process, Tahiri considered the correct calculation and checked the questions which most of his students could not answer correctly. In addition, Aziz did not look at his students' names and tried not to make any random mistakes. Deniz also used a program called Blackboard to grade her multiple-choice exams in some of her classes. Deniz and Tiger used different types of assessment methods in their classes. Deniz did not look at her students' names when she graded essays and open-ended questions, but Tiger looked at his students' names because he believed that he should take into account their attendance, participation and motivation in grading open-ended questions, which he considered as being fair in grading because he believed he gave his teacher evaluation grade in this way. Deniz graded her students' presentations inside and outside the class.

As found in the fourth observations, Beşiktaşlı, Black Eagle, Aziz, Tahiri, Crazy Soul and Aziz prepared their own grading tools by putting a transparent paper on a student bubble sheet on which they drew their answer keys to grade their students' exam papers. In addition to them, İlkbahar preferred preparing her grading tool by

cutting that paper. These participants put their grading tools on their students' bubble sheets, calculated the number of their correct answers and multiplied them with the points given to them. After calculating their students' grades, they wrote their students' grades on the students' exam papers, the excel sheets, or separate papers. Apart from them, Deniz and Tiger prepared open-ended questions. While grading open-ended questions, Deniz graded her students' papers separately. She cut a piece of paper, put it onto her students' names and paid attention to the use of the keywords related to the classroom discussions and readings and to conveying the message more than the use of grammar in grading her students' papers. Tiger also graded his students' papers separately, cut half point if the words were not written correctly and paid attention to the use of grammar because he believed that grammar was the thing that made what his students wrote understandable and to the connection between his students' answers and their classroom discussions in grading his students' papers. Both Tiger and Deniz wrote down the total points given to each question on their students' papers, summed them up and wrote their students' total grades on their papers.

4.3.3.7. Consistency of assessment interpretation

To make their assessment interpretations consistent, Beşiktaşlı and Tahiri told in the second interviews that they used their classroom observations. Beşiktaşlı matched his students' grades with their classroom performances, while Tahiri compared the average of his students' grades with the ideal average which he determined for one of his classes. Black Eagle and Tiger followed their assessment criteria strictly during the term and tried not to make any change in them, which Black Eagle considered as fairness. Crazy Soul and Black Eagle paid attention to objectivity in grading. Crazy Soul and İlkbahar cared to treat their students fairly and equally. In addition, Crazy Soul supported using the same type of exam format during the whole term. Besides, Deniz looked at the distribution of her students' grades. If she thought they were distributed normally, she thought that her grades were consistent. She also paid attention to the content validity of her exams like Aziz and to avoiding her students' cheating. Moreover, Aziz believed that if his exams had a relevant criterion, scoring and problem-free questions, he made the exams' results consistent.

4.3.3.8. Interpreting formal and informal student evaluation

The findings of the second interviews pointed out that Tahiri, Crazy Soul, Deniz and İlkbahar interpreted their formal assessment results based on the outcomes of whether their students learned what they taught and whether they became successful depending on their students' grades. Similarly, Beşiktaşlı interpreted his formal assessment results as whether his students passed or failed in his classes. In addition, Tiger found out his students' weaknesses and guided the ones who were willing to learn by interpreting his formal assessment results. Moreover, Tahiri, Deniz and Black Eagle self-assessed their exams and instruction by using their interpretations of their formal assessment results. Thus, they found out the problem(s) in their instruction and exams and tried to overcome them. Black Eagle and Crazy Soul also interpreted their assessment results as their success, that is, if their students got high grades in their exams, it meant to them that their students learned what they taught; therefore, they were satisfied with their students' results. Besides, Aziz believed that formal assessment results provided him with the proof about his students' learning. Tahiri, Deniz, İlkbahar and Black Eagle believed that informal assessment results revealed their students' real performances and learning. Accordingly, Black Eagle thought that such data enabled him to review what was not learned and Deniz said that such data determined the success of her instruction. However, Aziz claimed that informal assessment data indicated only the judgments of a teacher about his students.

As the focus group discussion revealed, each participant believed that selected response did not show their students' real performances. That is, they could not check whether their students could use what they learned in real life. Therefore, they could not interpret their assessment results like their students met their course objectives or their instructional methods became successful depending on the grades. Crazy Soul stressed out that if a student who did not study a lot got a high grade from an exam, he/she might have cheated; therefore, such situations should be taken into account in interpreting the assessment results. In addition, Black Eagle believed a teacher should self-assess his instruction based on his students' grades. Crazy Soul agreed with him and added that the teacher should assess himself as a teacher in terms of the content of his exams, the difficulty levels of the exams and the distribution of the questions in order to find out the source(s) of a problem related to the students' grades.

4.3.3.9. Using student assessment results for assessment tools and students' learning

As understood from the second interviews, Deniz, Beşiktaşlı, Tiger, Crazy Soul, Black Eagle and İlkbahar used their assessment data to improve their assessment tools by self-assessing their exams and instruction. Beşiktaşlı compared his students' grades with their classroom performances, while the rest used the data to find out the problems in their instruction and exams. If they encountered any problem, they tried to overcome them, so they could improve their assessment tools by using their assessment data. On the other hand, Tahiri and Aziz did not use the data for this purpose because Tahiri had a lot of students, lacked enough time and worked a lot and Aziz believed that he was not provided with the ideal teaching environment.

The second interviews also showed that Beşiktaşlı, İlkbahar and Crazy Soul could not use assessment results to find out their students' weaknesses because of the lack of time and the number of the students in their classes, while Aziz did not use the data for this purpose because he claimed that selected response did not reflect his students' real performances. On the other hand, Tiger, Tahiri, Black Eagle and Deniz used the assessment data to find out their students' weaknesses, yet Black Eagle and Tiger emphasized that they could help only the students who were willing to learn. Similarly, İlkbahar said she was able to use her assessment data with such students even though she mentioned that she could not use. Tahiri answered the questions which many of his students could not answer in his exams in his classes, while Deniz related her questions to the parts which her students did not pay enough attention to. She did so because she believed that this helped her students learn better. Last, Aziz used the observations and the technique "question-answer" to find out his students' weaknesses and strengths in the class to help these students.

4.3.3.10. Washback effect

As the second interviews indicated, Beşiktaşlı was aware of the washback effect of his exams, but he did not take into account this in his assessment and evaluation practices. Similarly, İlkbahar did not pay attention to the washback effect of her exams because she did not believe her exams had such an effect. On the other hand, the other

participants were aware of the fact that their exams had a washback effect on their students. Therefore, Tahiri made motivating speeches to his students in his classes, Tiger, Crazy Soul and Black Eagle prepared questions suitable for their students' levels of English and Aziz helped his students to understand that he did not use assessment as a punishment, but a tool to check their learning. In addition, Black Eagle paid attention to content validity, the types of the questions which his students were familiar with and his students' comments about the previous exams. Deniz also used her personal evaluation system about the difficulty levels of her exams, but this did not change the way she prepared her exams.

As revealed in the think-aloud protocol, the participants made their exams content valid in order to create a positive washback effect. They also prepared their questions by using the types which their students were familiar with. In addition, their questions reflected what they taught in their classes. As a result, they aligned their assessment tool with their course objectives.

In addition to the think-aloud protocol, the findings of the document analysis revealed that the participants used five stages for positive washback effects. They (a) made their exams content valid, (b) used the types of the questions similar to the classroom ones, (c) reflected what they taught with their questions, (d) aligned their exams with their course objectives and (e) gave enough time to their students to study for their exams.

4.3.3.11. Confidentiality of assessment and assessment results

During the second interview, Tahiri, Beşiktaşlı, Deniz, Black Eagle, Tiger, Crazy Soul and İlkbahar told that they kept their exam files locked in their offices in order to provide the confidentiality of their exams. Tiger carried his exams with him and İlkbahar saved her questions on her computer after she prepared them, so they made their exams confidential. Black Eagle did not share his questions with his colleagues apart from his coordinator and partner and he graded his exams on his own, which he thought made his exams and exams' results confidential. Besides them, Beşiktaşlı prepared and printed his exams out of class time, Tahiri printed his exams on the exam day and Deniz paid attention not to leave any document related to her exams in a place

where her students could easily reach. Moreover, Aziz took some precautions like preparing two types of booklets to avoid his students' cheating attempts, which he claimed made his exams more confidential. In addition, Aziz, Tahiri, Deniz, Crazy Soul, İlkbahar and Black Eagle announced their grades on the online information system of the university which could be used only by them, so they believed that they made their exam results confidential. Deniz and İlkbahar also gave their exam files to the related departments after their exams were over.

The findings of the think-aloud protocol indicated that the participants cared the confidentiality of the exams. Tahiri and Beşiktaşlı prepared their exams at home on their computers, while Crazy Soul, Aziz and İlkbahar prepared their exams in their offices on their computers. In addition, Black Eagle and Tiger prepared their questions on papers in their offices out of their class times and locked their papers in their drawers. Deniz chose her questions in the copies of the test booklet in her office out of class time and locked the copies in her drawer.

In addition to the think-aloud protocol, the fourth observations revealed that the participants gave importance to the confidentiality of their exams and exam results. To make their exams confidential, Beşiktaşlı, Crazy Soul and Aziz recorded their exam results on the excel sheets, while the others recorded the results on the online information system. Last, Beşiktaşlı, Deniz and Black Eagle kept their exams and exam results locked in their drawers in their offices, while the others kept them locked in their cabinets in their offices.

4.3.3.12. Attitude toward exam complaint

As stated in the second interviews, the attitudes of the participants toward their students' exam complaints were generally shaped by the types of the exams (Beşiktaşlı, Deniz and Aziz), their students' attitudes (Crazy Soul, Black Eagle and Tiger) and the number of the students (Tahiri and Crazy Soul). Beşiktaşlı distributed his quizzes to his students and checked the exams with them in his classes, wanted his students to come to his office to check their midterm exams and wanted them to write a petition if they wanted to reject to their grades in the final exam. Similarly, Aziz wanted his students to write a petition to check their midterm and final exams because their percentages in his

grading system were 90%, but his students could come to his office and check their quizzes without writing a petition because the percentage of the quiz was 10%. Like Beşiktaşlı, Deniz wanted her students to come and check their quizzes and midterm exams in her office, while they had to write a petition for their final exams. Secondly, Black Eagle allowed some students who he knew studied a lot, but could not get a high grade and did not make a complaint about every exam to come to his office in order to check their papers, but he wanted the rest to write a petition to their departments for this. Crazy Soul also checked the exam papers with such students in his office, but wanted the rest to write a petition in order to check their papers. Besides, if his students came to his office with a good attitude to check their papers, Tiger allowed them to check their exams. Thirdly, if the number of the students were low, Tahiri and Crazy Soul allowed them to come to their offices and check their exam papers, but if it was a lot, they wanted their students to write a petition to their departments. Apart from these approaches to the students' exam complaints, İlkbahar's students could check their exams with her in her office or send an e-mail to her about checking their exams again.

4.3.4. Using assessment results in making decisions about student, instruction, school and curriculum

This part details the analysis of the fourth sub-component of LAL based on the codes in Table 4.10.

Table 4.10.

Codes of the Fourth Sub-component of LAL

4. Using assessment results in making decision about student, instruction, school, and curriculum
4.1. Evaluating assessment data
4.2. Wrong and correct evaluation of assessment data
4.3. Developing instructional plan for students
4.4. Making changes in instruction and curriculum

This sub-heading details the issues in Table 4.10.

4.3.4.1. Evaluating assessment data

The findings of the second interviews showed that Aziz evaluated his assessment data in terms of whether his students became successful by comparing their grades with the grading system of the school. Likewise, if more than half of his students got high grades from his exam, Beşiktaşlı believed “I could teach my students well. My students learned well and reflected what they learned in their performances.” Yet, Tahiri evaluated his assessment data “in terms of overcoming my students’ and my weaknesses and of teaching again what my students did not learn.” Tiger believed that assessment data should be evaluated “in terms of how students can learn better in the next steps and with which method they can learn better.” In addition to these ideas, Crazy Soul evaluated his assessment data “to see a student’s situation in our classes, to evaluate the assessment tool’s situation, to understand to what extent we have taught as teachers and whether we need to improve ourselves.” Similarly, Black Eagle evaluated his assessment data as “there must be a problem either in my teaching style, in my students’ studying style, or in the exam.” İlkbahar and Deniz also evaluated their assessment data in terms of self-assessing their teaching, exams and course materials.

As seen in the second interviews, the participants evaluated their assessment data as checking what their students had and had not learned in their courses. They also made an evaluation of their assessment data in terms of self-assessing their instruction, goals and assessment tools. In accordance with this, their evaluation of their assessment data included checking whether their instruction became successful depending on their students’ grades.

4.3.4.2. Wrong and correct interpretation of assessment data

The second interviews indicated that Beşiktaşlı, Tahiri, Black Eagle and Crazy Soul thought that the correct evaluation of their assessment data was related to self-assessing their teaching, goals and assessment tools. According to this, Tahiri, Crazy Soul and Black Eagle related the correct evaluation of their assessment data to taking responsibility in making changes in their teaching, goals and assessment tools depending on the results of their self-assessment. In addition, Crazy Soul believed that basing the evaluation of the data only on the students’ grades was wrong, so he

supported to take into account the students' classroom performances. Like him, Deniz claimed that considering a teacher and his students as successful and unsuccessful depending on the students' grades was the wrong evaluation of the assessment data. Besides, Tiger connected the correct and wrong evaluation of the assessment data to whether a teacher followed his assessment criteria strictly. In terms of the wrong evaluation of the data, Beşiktaşlı believed that the data obtained from an invalid assessment tool caused the wrong evaluation, while he connected the correct evaluation of the assessment with the 70% success, which meant the 70% of his students had high grades in one of his classes. Like Beşiktaşlı, İlkbahar thought that always considering the data obtained as correct led to the wrong evaluation of the data. In addition, Aziz related the correct evaluation of the assessment data to sharing the data with other colleagues and discussing the data together by being aware of the effects of the chosen assessment method.

4.3.4.3. Developing instructional plan for students

As understood from the second interviews, Beşiktaşlı, Tahiri, Deniz, Black Eagle and Tiger used their assessment data to develop instructional plans for their students by self-assessing their teaching, goals, the content of their classes and their assessment tools. Accordingly, Tahiri reviewed what he considered as important with his students, realized his mistakes and tried not to do them again, while Black Eagle, Tiger and Deniz decided whether they should go on using the same teaching method or should change them depending on their students' grades. On the other hand, Crazy Soul, İlkbahar and Aziz could not make such plans due to the lack of time, syllabus, the difference between reality and idealism, the number of the students in their classes, their workload and not believing in the efficiency of the chosen assessment method.

4.3.4.4. Making changes in instruction and curriculum

As the second interviews indicated, Beşiktaşlı, Tahiri, Deniz, Black Eagle, Tiger, Crazy Soul and İlkbahar made some changes in either their instruction or curriculum, or both. They self-assessed their teaching, goals and the content of their lessons. As a result of this self-assessment, Beşiktaşlı, Tahiri, Deniz, Black Eagle, Tiger

and Crazy Soul made some changes like increasing or decreasing the weights of some parts of the content in their teaching, but not in their curriculums, but Deniz and İlkbahar told that they made changes in the courses for which they prepared the curriculum on their own. İlkbahar told she did not make any change in her teaching method as she decided it according to the topics, while Deniz added some productive activities to the syllabi prepared by the others and given to her. However, Aziz told he did not make any change in his curriculum and teaching method owing to the number of the students in his classes, his workload and not having a suitable language teaching environment.

In addition, Black Eagle stated in the focus group discussion that all participants used their assessment results in evaluating their course books which they used in certain departments. He added that their evaluation was based on their students' grades and they either changed their course books and prepared a new curriculum for their new course books or went on using the same course books with the same curriculums. In addition, İlkbahar emphasized their assessment results indicated them whether their students benefitted from their course materials. Besides, Beşiktaşlı told they increased the number of the activities related to a part which their students were not good at in their courses which they found out through their students' grades in order to help their students to improve themselves in that part of their courses. The rest agreed with these participants' opinions about using the assessment results in decision making about their students, curriculum and instruction.

4.3.5. Developing valid grading procedures using students' assessments

This part informs about the analysis of the fifth sub-component of LAL depending on the codes in Table 4.11.

Table 4.11.

Codes of the Fifth Sub-component of LAL

5. Developing valid grading procedures using students' assessment
5.1. Grading systems
5.2. The reasons for using the grading systems
5.3. Developing grading systems
5.4. Validity of the grades given through the grading systems
5.5. Purposes of giving grades

This part presents the issues in Table 4.11 in a detailed way.

4.3.5.1. Grading systems

As the second interviews indicated, Tahiri, Deniz, Black Eagle, Tiger, Aziz, İlkbahar and Crazy Soul used one quiz, midterm and final exam in their grading system which was used for academic English classes. The quiz constituted the 10% of the total grade, midterm the 40% of the total grade and final the 50% of the total grade. While Beşiktaşlı used one midterm and final exam with the same percentages, he used two quizzes instead of one quiz and determined that each quiz constituted the 5% of the total grade. Like him, Crazy Soul omitted one quiz in his academic English classes with the students who had English preparatory school education and added a presentation, instead. On the other hand, Deniz developed her own grading system for one of her courses: communicative competence course. It was composed of one midterm, presentation and one quiz. Her midterm exam made up the 30% of the total grade, presentation the 50% of the total grade and quiz the 20% of the total grade. In her advanced reading class, she had one midterm and one final exam in her grading system and their percentages in the total grade were 40% and 60% in order. In the advanced writing class, Deniz had one quiz, one final exam and one portfolio and their percentages in the grading system were respectively 10%, 50% and 40%.

4.3.5.2. The reasons for using the grading systems

During the second interviews, Tiger, Crazy Soul and Beşiktaşlı told that they used midterm and final in their grading system because of the legal obligation. Yet, Tahiri told that he wanted to vary his assessment tools, encourage his students to study and give them enough time to study for his exams. He also thought that midterm and final exams were generally used, so he had them in his grading system. Like him, Deniz, Black Eagle, Aziz, İlkbahar and Crazy Soul talked about a different reason for using one midterm exam, final exam and quiz in their grading system. They told they had to have these exams in their grading systems because of the mutual decision on the grading systems in their academic English classes. The below excerpt illustrates this.

Crazy Soul (I2): Each component of the system was determined based on a mutual decision. We [participants] talked and discussed the things like how many midterm exams we should use, how many presentations our students should make and whether we should give an assignment in our department meeting.

Yet, Beşiktaşlı and Tiger had two different personal reasons for using quiz in their grading system. Beşiktaşlı used one quiz in order to enable his students to find out their mistakes and to avoid their anxiety as the quotation below clearly shows.

Beşiktaşlı (I2): ... because our [participants'] aim is to teach vocabulary related to our students' departments to them, we decided to make two quizzes related to vocabulary teaching: one quiz before and one quiz after the midterm exam in order to reduce our students' exam anxiety and to enable them to overcome their weaknesses.

However, Tiger used one quiz because he considered it as a requirement for language learning and made his students be ready for his other exams. To indicate:

Tiger (I2): We [participants] do not have to make one quiz. We make it because we want our students to be always ready for an exam and think that it is a part of foreign language education. This is why we make it.

In addition to these reasons, Crazy Soul added that he had to include one quiz, midterm exam and final exam in his grading system because of the lack of time and his syllabus. However, Crazy Soul replaced one quiz with presentation in one of his academic English classes because he believed that he had a right to take the initiative as a result of his teaching experience and teaching context. He also wanted his students to use English outside the class, improve their public speaking and understand that English is used in their major. The quotation below clearly supports this.

Crazy Soul (I2): Of course, we [participants] can take our own initiative. Of course, as a teacher, we can make small changes in the system depending on our experiences, our students' situations and our conditions. In fact, I developed my own system to some extent in academic English. We have one midterm exam and final exam, but I have replaced quiz with presentation. I have not decided the percentage of the presentation in the grading system. It may be more than 10%. It may be 15% or 20% depending on my students' situations. I give my own decision about this in the following way: I can have opportunities to increase the percentage because I assess many things in my students' presentations like their research skills and public speaking.

On the other hand, Deniz developed her own grading system for her communicative competence course depending on the content of the course. She also

made changes in the grading systems of her elective courses which were designed by her colleagues if she found something illogical in those grading systems.

4.3.5.3. Developing grading systems

As indicated in the second interviews, all participants joined the meeting to develop the grading system for academic English course. In this meeting, according to Tahiri, the components of the grading system were chosen depending on the experiences of the previous course teachers as the quotation below demonstrates.

Tahiri (I2): I participated into the development process. We [participants] made a meeting together and developed this system through our discussions. That is, it was developed mutually. As I said, one or two meetings had been made about this issue before. In addition, we have a three- or four-year experience in the department. Besides, how can we assess our students best and optimize our workload? These were taken into consideration and the system was developed in this way.

İlkbahar also mentioned that they shared their ideas with each other about the grading system and used the most voted one as their grading system, but Tiger claimed that he expressed his ideas, but his ideas were ignored. In addition, Crazy Soul and Beşiktaşlı said that the participants used one midterm and final exam because of the legal obligation; Tahiri and Aziz claimed that this grading system was the best to assess their students and optimize their workload depending on their work conditions; İlkbahar told they used midterm and final exams to maintain their instruction. In terms of the quiz, Beşiktaşlı told that the participants decided to use it because their students could find out their weaknesses, overcome them and get high grades. İlkbahar added that the participants used one quiz to make their students be ready for midterm and final exams.

Black Eagle, Crazy Soul and İlkbahar told in the second interviews that they determined the percentages of each component in the academic English grading system according to the content coverage of the exams as the quotations of Black Eagle and İlkbahar clearly indicate.

Black Eagle (I2): How did we [participants] determine it? First, everybody was asked to explain what they thought about the system. Then, it was decided that this grading system is the best one according to everybody's opinions. We make our quiz before the midterm exam as a mini exam to help our students be ready for the midterm exam. Besides, it covers only what is taught within three or four weeks. Therefore, its percentage must be low. The final exam covers almost what

we teach during the whole term, so its average must be the highest, 50%. Then, we had 40%. We gave it to the midterm exam.

İlkbahar (I2): As the final exam covers from the beginning to the end of the term and as we [participants] consider it as a general evaluation, it has the highest percentage. The midterm exam is an important exam like the final exam, but not as important as it, so it has a percentage close to the final exam's percentage. The quiz is used in order to prepare our students for other exams and to maintain their learning dynamic, so it has the lowest percentage. We all thought it should be like this in the department. It was not my decision. It is a mutual decision.

Yet, Beşiktaşlı said the participants gave these percentages to their midterm and final exams in order to make their students study harder and the percentage for their quiz in order to help them realize and overcome their weaknesses.

Apart from the other participants, Deniz developed her grading system based on the content of her lesson and determined the percentages depending on the weights of the activities in her communicative competence course. To illustrate:

Deniz (I2): I gave 30% to my midterm exam by calculating the number of reading passages in communicative competence II. I have nine reading passages in total. We [the participant and her students] have studied one passage each week and the number has become six in six weeks. This number made me give 30% to my midterm in my grading system. Then, there are three more reading passages left. My quiz will be about them, so I gave 20% to it in my grading system. Though the percentages of the midterm and quiz change in my mind, I do not want to make changes in my syllabus in order to be consistent. My classes generally end with discussions, so they include speaking. I want my students to make their talks in a more specific format. As we are going to make presentations during the last three weeks, I gave 50% to speaking in my grading system. We [the participant and her students] are learning what to do when we speak. We are learning them for a purpose. Besides, I wanted them to produce something. That is why I gave this percentage to speaking in my system.

In addition, she made some changes in the grading systems of her elective courses because she did not find some parts logical. Thus, she omitted them and redistributed the percentages. To indicate:

Deniz (I2): If it [the grading system] sounds logical, I do not make any complaint, but if it does not sound logical and fair, I question it. If I can discuss it with my administration, I discuss. If my words are paid attention, we do the necessary changes.

The participants also explained in the focus group discussion that they decided to use midterm and final exams as they were generally used in assessment and evaluation at universities. They also used the quiz to make their students study for their

lessons, find out their weaknesses and be ready for their other exams. They emphasized that the percentages of their quiz, midterm and final exams were determined based on their content coverage.

The participants made two meetings to choose their assessment method for the term and the researcher observed them in these meetings. In the first observation, they wanted to go on using selected response in the second term because Crazy Soul, İlkbahar, Beşiktaşlı and Tahiri shared their experiences with the new members of the department (Deniz, Tiger, Black Eagle and Aziz). According to the old members, their students had copied and pasted information from Wikipedia without doing their own research. Then, they had used this information in preparing their presentations and writing their reports without changing. Their reports had not been prepared in the expected format and their presentations had not been made on the scheduled days. However, they had made complaints about their presentation grades after their announcements. According to their students, they had not deserved those grades because they had spent a lot of time on preparing them. Consequently, watching their students' presentations, giving grades to them, checking who had made the presentations on time, controlling who had submitted the reports on time, reading their reports, grading the reports and dealing with their complaints about their grades led to an increase in their workload. In addition, they added that their students had not followed the presentation rules though the participants had explained the rules to them many times in their classes. Furthermore, some of the new participants added that they wanted to give their students a teacher evaluation grade because some students studied a lot, but their grades were low and they failed. Therefore, they wanted to help such students by giving an evaluation grade. Nevertheless, the old members told that they had given such a grade to their students in the previous terms, but their students had complained about their grades by making such comments as "Why did you give this grade to me?", "I came to all of your classes, but my evaluation grade is low, why?" and "Student X did not come to your class regularly, but you gave him a higher grade than me, why?". Even though they had been explained how that grade had been given to each student, they had not taken into consideration in their complaints. The old participants told their workload had increased due to these complaints. As a consequence, the old participants had given up giving teacher evaluation grades. They

also added that selected response avoided such problems. It enabled them to grade their students' papers objectively and give the grades which they deserved depending on their knowledge. The participants decided to use selected response in order to reduce their workload because of the number of the students in their classes, high expectations from them and the lack of time. Then, they decided to use one midterm, one final exam and two quizzes in the first meeting because they had to make one midterm and one final exam due to the legal issues and wanted to make two quizzes to make their students study, find out and overcome their weaknesses and be ready for the other exams. They gave the highest percentage (50%) to the final exam as it covered the units studied during the term, the second highest percentage (40%) to the midterm exam because it covered the units studied during the first eight weeks and the lowest percentage (10%) to two quizzes by sharing 10% as 5% for each quiz. However, they had another meeting about the grading system. In the second meeting where the second observation was made, some participants wanted to make one quiz in this term because they did not have enough time to prepare two quizzes, to grade them and to announce their results. The others agreed with those participants' suggestions as they complained about the number of the students in their classes, the number of the classes they had to teach and the different courses they had to make preparation for before teaching them. As a result, they reduced the number of the quizzes to one and gave the total percentages of two quizzes (10%) to one quiz.

4.3.5.4. Validity of the grades given through the grading systems

In terms of the validity given through their grading systems, almost all participants thought that their grades were valid as the second interviews indicated. However, Beşiktaşlı did not believe that the grades were not valid as the grades obtained from selected response did not show his students' real performances and Aziz believed that his grades had low validity because he and his colleagues used the same exams at different times, so their students might cheat. The grades were valid for Tahiri and İlkbahar because of deciding the percentages of the midterm and final exams in the grading system based on the exams' content coverage and of the quiz with the aim of engaging their students in learning, maintaining their instruction and making their students ready for the other exams. They were valid for Black Eagle since everything

was student-centered, he had certain assessment criteria and he took the necessary precautions during the whole assessment process. Crazy Soul found his grades valid as the system was product-oriented, so it reflected what his students learned. Black Eagle and Crazy Soul also added that the system provided them with objective grading, which they thought made their grades valid. Tiger considered his grades as valid because he prepared his questions suitable to his students' levels of English and tried to be fair in grading. Deniz also told that she always explained her assessment criteria and system to her students, which she believed made her grades valid.

4.3.5.5. Purposes for giving grades

The second interviews showed that all participants used their assessment results to decide who passed and failed depending on their students' grades. In addition, Deniz and Tiger used their assessment results to self-assess their teaching.

4.3.6. Communicating assessment results to students and other stakeholders

This part informs about the analysis of the sixth sub-component of LAL made based on the codes in Table 4.12.

Table 4.12.

Codes of the Sixth Sub-component of LAL

6. Communicating assessment results to students and other stakeholders
6.1. Communicating assessment results to students and school administration
6.2. Meanings of assessment results for different stakeholder
6.3. Correct interpretations of the results
6.4. Limitations in interpreting assessment results
6.5. Reflections of interpreting assessment results
6.6. Avoiding misinterpretation of assessment results
6.7. Avoiding possible measurement errors in communicating assessment results

In this sub-heading, the issues in the table 4.12 will be explained in detail.

4.3.6.1. Communicating assessment results to students and school administration

From the findings of the second interviews it was found that all participants announced their assessment results to their students on the online information system of the school. Beşiktaşlı, Tahiri, Deniz, Black Eagle, Aziz and Crazy Soul took grade reports from the same system, signed them and gave them to their administrators. Tiger told that his administrators could see the grades on the same system, while İlkbahar believed her coordinator communicated their assessment results to her administrators.

In addition, the focus group discussion showed the preferences of each participant in communicating their assessment results to their students. Beşiktaşlı told he used his wiki page to announce his exam results regularly, so his students got used to it. Like him, Crazy Soul said he printed out his results and stuck them on his office door. On the other hand, the other participants communicated their assessment results through the university's online information system.

As the fourth observations revealed, Tahiri, Deniz, Black Eagle, Tiger, Aziz and İlkbahar entered their grading criteria on the online information system by using their own usernames and passwords before announcing their students' grades. After that, they wrote down their students' grades on the system. They compared the grades on the system with the ones on their papers. If there was no difference, they saved and announced the grades on the system. Though Beşiktaşlı and Crazy Soul told they would communicate their exam results with their students by using the same online system, they communicated their results in different ways. Beşiktaşlı converted his students' grades on the excel sheet to a pdf file by showing their ID numbers and grades on the pdf file and uploaded it to his wiki page. In addition, he printed out one copy of that page and stuck it on the door of the class whose papers he graded before his lesson started. Meanwhile, Crazy Soul converted his students' grades on the sheet to a pdf file which included their names, ID numbers and grades, printed out the file and stuck it on his office door.

4.3.6.2. Meanings of assessment results to different stakeholders

In terms of the meanings of their assessment results to their students and administrators, Beşiktaşlı and İlkbahar thought that their results meant success to their students and administrators as understood from the second interviews. İlkbahar also added that they meant happiness or sadness to her students and administrators depending on the grades. Tahiri and Deniz believed that the meanings of their assessment results to their students was pass and fail. Accordingly, Tahiri thought that they also meant learning what was taught to his students. In accordance with these, Tiger, Crazy Soul and Aziz claimed that self-assessment (knowing their levels, strengths and weaknesses) was what the grades meant to their students. According to Black Eagle, high grades meant “I deserve it” to his students, whereas low grades meant blaming their teachers for their failure to them. According to Crazy Soul and Aziz, the meaning of their assessment results was teacher evaluation for their administrators; Tiger and Deniz thought it was program evaluation for their administrators; Tahiri thought it was student satisfaction for his administrators.

4.3.6.3. Correct interpretations of the results

As told in the second interviews, Beşiktaşlı, İlkbahar and Aziz supported that the correct interpretations of their assessment results must be made based on their students’ real performances. Tiger and Deniz believed that the correct interpretation should be related to their students’ self-assessment, that is, the fact that their students should find out their weaknesses and try to overcome them. Tahiri claimed that pass and fail were the correct interpretations for his students if they interpreted their grades based on what they learned and did not learn. In addition, Tiger and Black Eagle supported that the correct interpretation should be made depending on the objective grading. Black Eagle also believed that if everything was student-centered, transparent and required his students to take responsibility in their learning, the assessment data could be interpreted correctly. Besides, Crazy Soul and Tahiri mentioned that teacher evaluation should be made based on the interpretations of students’ real learning performances.

4.3.6.4. Limitations in interpreting assessment data

In the second interviews, Beşiktaşlı, İlkbahar, Crazy Soul and Aziz told that they considered selected response as a limitation in interpreting their assessment results because selected response did not show their students' real performances and their students could answer such questions without knowing the answers. Beşiktaşlı believed that not having a right to make changes in the grading system was also a limitation. According to Deniz, that her students did not know their capacities, that they had high expectations, that they were not competent in a skill and that they had some behavioral problems limited the interpretations of her assessment results. Besides, not following the grading system strictly and students' egocentrism were the limitations for Black Eagle; being compared with other teachers because of the chosen assessment methods was a restriction for Tiger; making an evaluation based only on grades limited the interpretation of his assessment data for Crazy Soul.

4.3.6.5. Reflections of interpreting assessment data

Beşiktaşlı, Deniz, Crazy Soul and İlkbahar said in the second interviews that questioning the reason for failure was the most common reflection of interpreting their assessment results. As a result, Deniz and Crazy Soul claimed that students might blame their teachers for their failure, while Deniz, İlkbahar and Beşiktaşlı told they might realize their weaknesses and try to overcome them. Accordingly, Tiger claimed that they might want a teacher to change his assessment method(s). According to Aziz, they might generalize their assessment results. As a result, Black Eagle thought that teachers had to repeat their assessment criteria and grading systems many times to their students. Besides, Aziz claimed that the administrators might have wrong expectations from their students and teachers. In addition, Tahiri and İlkbahar thought the reflection of their results' interpretations based on grades was a success to both students and teachers, while Crazy Soul believed that the reflection was being considered successful for teachers.

4.3.6.6. Avoiding misinterpretation of assessment data

Beşiktaşlı, Black Eagle, Tiger, Crazy Soul, Aziz and İlkbahar believed that their students and administrators should be informed about the assessment criteria and grading system in order to avoid the misinterpretation of their assessment results as understood from the second interviews. In accordance with this, Tiger mentioned that their students should be taught about how to think about their assessment results, to find out their weaknesses and to overcome them, while Deniz believed that teachers should not use performance and constructed response assessment methods in their assessment systems, but they should ask memorization questions. Yet, Aziz and Beşiktaşlı told they did not inform their administrators about their assessment criteria and grading systems because the administrators did not want them to do so. In relation with this, Crazy Soul and İlkbahar supported that teachers and students should join the meetings with the administrators in which their assessment results were evaluated. Crazy Soul also stated that teachers should self-assess their assessment tools for avoiding the misinterpretations of their assessment data.

4.3.6.7. Avoiding possible measurement errors in communicating assessment results

Deniz, Tiger and Aziz double checked their exams after they prepared the exams in order to avoid possible measurement errors as found in the second interviews. Deniz believed that she avoided inconsistencies in her questions by doing so. Like them, Black Eagle told he prepared his questions very carefully and prepared his answer keys after preparing his questions. In addition, Tiger and Tahiri depended on peer evaluation. Tahiri and Aziz canceled the problematic questions and redistributed the points. Similarly, Crazy Soul informed his students as soon as he noticed the mistakes in his questions. Crazy Soul also tried to make his grading objective by using selected response. Moreover, İlkbahar compared her students' grades with their classroom performances, double checked the ones whose grades were low, but were successful in her classes and used technology to announce her grades to avoid possible measurement errors. Besides, Beşiktaşlı checked his quizzes with his students and checked his

midterm and final exams on his own by comparing the grades on his excel sheet and his students' exam papers.

In accordance with the previous paragraph, the participants were observed to double-check their answer keys by answering their questions again, preparing new answer keys if necessary and comparing the new answer keys with the old ones to avoid any possible measurement errors in the fourth observations. Beşiktaşlı, Black Eagle, Tahiri, Crazy Soul, İlkbahar and Aziz double-counted the number of their students' correct answers while grading. Tiger and Deniz double-checked the grades they gave to their students' answers to the open-ended questions in their exams in order to be sure that they followed their own criteria in grading open-ended questions. The participants double-calculated their students' total grades. They also compared the grades on the online information systems or the excel files with the ones on their students' exam papers before announcing them.

4.3.7. Recognizing unethical, illegal and inappropriate assessments and uses of assessment information

This part presents the analysis of the seven sub-component of LAL depending on the codes in Table 4.13.

Table 4.13.

Codes of the Seventh Sub-component of LAL

7. Recognizing unethical, illegal and inappropriate assessment methods and uses of assessment information
7.1. Ethical, legal and professional assessment and evaluation practices
7.2. Problems encountered and how they were overcome

This sub-heading details the issues in the Table 4.13.

4.3.7.1. The ethical, legal and professional assessment and evaluation practices

During the second interviews, Beşiktaşlı, Deniz, Tiger, Crazy Soul and İlkbahar told that being fair and objective in grading is an ethical, professional and legal behavior that a teacher should follow. According to them, it includes grading his students' papers

correctly, not being affected by the ideas about his students, not giving extra information to his favorite students, not helping his students during the exam and not making a change in their grades. In addition, Tahiri, Black Eagle, Crazy Soul and Aziz believed that asking questions related to what he teaches in his classes is another behavior that the teacher should follow. Similarly, Tiger considered preparing questions relevant to the students' cognitive levels as another ethical and professional behavior that the teacher should have. Black Eagle, Aziz and İlkbahar also added that informing his students about his course goals and assessment criteria is the third behavior that the teacher should adopt. Accordingly, Aziz believed that the teacher should prepare his course materials based on his course goals and give enough importance to teaching the materials. In terms of the assessment criteria, Crazy Soul, Tiger and İlkbahar insisted that the teacher should not make any change in his assessment criteria, but should follow the criteria strictly. Besides these behaviors, Tahiri and Black Eagle believed that the teacher should announce their assessment results timely, while Beşiktaşlı and Crazy Soul emphasized that the teacher should take into account the copyright issues. Tiger individually suggested that the teacher should choose assessment methods which his students could do, while Aziz individually supported that the teacher should be ready to help his students based on their assessment results.

In addition, the focus group discussion pointed out that being fair and objective in grading was one of the legal, ethical and professional behaviors that the teacher should follow in assessing and evaluating his students. They added that not sharing questions with the students before the exam, providing the confidentiality of the exams and exam results, avoiding cheating, asking questions related to what was taught and being consistent in the difficulty levels of the exams were the other ethical, legal and professional teacher assessment and evaluation behaviors.

The participants determined eight ethical, legal and professional behaviors that the teacher should follow in his assessment and evaluation. These are (1) being fair and objective in grading, (2) not helping some students during the exams, (3) not making any change in grading, (4) grading the students' papers correctly, (5) providing the content validity of the exams, (6) making the exams and exam results confidential, (7) announcing grades timely and (8) paying attention to the copyright issues.

Determining such behaviors also required to check whether the determiners followed their own explanations. The participants were checked for this purpose with different data collection tools. The data collection tools revealed that almost all participants followed their criteria. The think-aloud protocol showed some participants did not pay attention to the copyright issues. The think-aloud protocol and document analysis indicated the participants made their exams content valid. During the fourth observations, the researcher observed that the participants graded their students' exam papers cautiously. They tried to be fair and objective in grading. They did not make any change on behalf of any of their students. They announced their students' grades timely. During the third observations, the researcher observed that none of the participants answered some students' questions because their questions were indirectly related to the answers to some questions. In addition, the think-aloud protocol and third and fourth observations demonstrated that the participants tried to make their exams and exam results confidential through some ways.

4.3.7.2. The problems encountered and how they were overcome

In terms of the problems related to these ethical, legal and professional behaviors, Beşiktaşlı, Black Eagle, Deniz, Tiger, Crazy Soul, Aziz and İlkbahar encountered some problems as clearly stated in the second interviews. Beşiktaşlı sometimes thought that he could not be objective in grading some students' exams, but he told that he questioned his attitudes toward those students, so he could avoid this problem. Deniz and Black Eagle had some problems because their students might ask them to make changes in their grades directly, or they might complain about their exams and make comments about their exams in order to have them change their grades indirectly. However, Deniz explained her assessment criteria again and wanted her students to empathize her, while Black Eagle carefully checked his exams by self-assessing his questions and followed his assessment criteria strictly in order to avoid such problems. Tiger faced some problems because of his colleagues who did not take assessment and evaluation seriously. He did not do anything to stop such problems, but he took lessons from such things and tried not to do them in his own assessment and evaluation. İlkbahar and Crazy Soul encountered cheating problems because of the other teachers who proctored their exams. Those teachers dealt with their cell phones

and did not proctor well. İlkbahar had problems because of this, so she tried not to do this in some exams where she was the only one who proctored and she reported the students who cheated. Crazy Soul found another proctor, prepared two types of booklets, arranged his students' seating and collected their cell phones to avoid cheating in some of his exams in which he could not proctor because of proctoring his other exams at the same time. Lastly, Aziz told that he might sometimes prepare a question which was not related to what he taught. In this case, he canceled the question and gave its points to his students. In addition, he prepared some of his exams with one of his colleagues, so each colleague might sometimes prepare questions related to something which one of them did not teach. In this case, he and his colleague informed each other about what they taught and did not teach in their classes before preparing their exams.

4.4. Conclusion

This chapter has presented the findings related to the implementation of the seven sub-components of language assessment literacy by describing the findings thickly. The next chapter is about the discussion of the findings according to the research questions by relating the findings to the literature review.

CHAPTER FIVE

5. DISCUSSION

5.1. Introduction

This part discusses the findings of the study by relating them to the related literature. It presents the discussion by following the order of the research questions.

5.2. Implementing the Sub-components of Language Assessment Literacy in the Class

The seven standards of assessment literacy put forward by AFT and its partner organizations (1990) were adopted as the sub-components of language assessment literacy in this study because these standards were used to structure the studies on assessment literacy and language assessment literacy. Consequently, this research question is discussed under seven sub-titles.

5.2.1. Choosing appropriate assessment methods for instructional purposes

Choosing appropriate assessment methods for the instructional purposes is the first step and sub-component of language assessment literacy. To choose assessment methods, it is important for a teacher to know the purpose, the intended test takers, and content and skills to be tested (JCTP, 2002), which the Turkish EFL instructors in this study do not consider more while choosing their assessment methods. Instead, they consider experience more, which is highly stated in the relevant literature (Scarino, 2013; Shohamy et al., 2008; Yildirim, 2012). As Davison (2004) emphasized, teachers gain negative experiences in using assessment methods because of their students. Consistent with the finding of Davison, the EFL instructors have negative experiences because of their students. Therefore, they are forced to select a certain type of assessment method: selected response which minimizes such problems in the EFL instructors' assessment.

The instructors do not pay attention to many issues including the purpose and intended test takers which are stated in the literature (MAC, 2013). According to Brown (2004) and MAC (2013), the language teacher should also be familiar with the purposes of assessment which are proficiency, placement, achievement and diagnostics. The instructors in this study are familiar with the two of the assessment purposes: achievement and diagnostics. Herrera and Macias (2015) stated that language assessment can be used to check what students have learned, decide who passes and fails, self-assess instruction and improve students' learning. Similarly, the instructors use language assessment to decide who passes and fails. They also utilize assessment for diagnostic purposes by self-assessing and improving their teaching as well as helping their students to find out and overcome their weaknesses in their learning.

Besides, the teacher should have enough knowledge about different assessment methods and their strengths and weaknesses in choosing assessment methods (AFT et al., 1990). It seems that the instructors have sufficient knowledge about assessment methods, but they acquire this knowledge mostly through their experiences. As stated in both national and international studies (e.g. Hatipoğlu, 2010, 2015a; Hatipoğlu & Erçetin, 2016; Lam, 2015; Kahl, Hofman, & Bryant, 2013; Koh & Velayutham, n.d.; Popham, 2006; Stiggins, 1995; Webb, 2002), most of the instructors mainly think that their pre-service/pedagogical formation assessment training ineffective and inefficient. Such experiences can be considered to have helped them to learn the strengths and weaknesses of different assessment methods. As a result, the instructors determine the quality of their exams depending on their experiences. The instructors are aware of the strengths and weaknesses of different assessment methods, which has forced them to choose and use selected response. They believe that selected response avoids several problems that performance assessment has caused in their assessment and instruction. Therefore, most of them believe that it increases the quality of their exams and instruction. As MAC (2013) stressed out, the effects of quality assessment increase effective teaching.

In addition, the teacher needs to know and understand the criteria used for choosing and evaluating assessment methods according to instructional plans, which enables him to consider different factors in selecting and evaluating the available assessment methods (AFT et al., 1990). Yet, the instructors develop their own criteria to

choose and evaluate assessment methods based on their experiences. Experience has caused them to prioritize reality, but not instructional purposes because the instructors think that their purposes are idealistic, but not compatible with reality. Different studies in the literature (e.g. Gökçe, 2014; Gönen & Akbarov, 2015; Lam, 2015; Örsdemir, 2015; Shohamy et al., 2008) support this finding as the inconsistency between idealism and reality causes teachers to ignore idealism and to focus on reality.

Another issue is the instructors' negative experiences related to performance assessment. Thus, the instructors tend to choose and use selected response more. Using selected response makes the instructors believe that it can be administered easily, that it is useful as it reduces their workload and that it is fair since it measures only what their students know. However, they know that it is not technically adequate to understand their students' real performances. Therefore, they only consider administrative appropriateness, usefulness and fairness in this step to be aware of the effects of valid and invalid assessment data on their instruction and students' learning because according to AFT and its partner organizations (1990), these three issues help to understand, learn and know the effects of valid and invalid assessment data on instruction and students' learning. In addition, Mertler (2003), White (2009) and Witte (2010) stated that assessment data is the key determiner of the quality and effectiveness of instruction and students' learning, so valid and invalid assessment data affects instruction and students' learning differently. The instructors are aware of the effects of valid and invalid assessment data because most of them use assessment data for improving their instruction by self-assessing it and for improving their students' learning by helping them find out and overcome their weaknesses.

In addition, the teacher should also have sufficient technical knowledge to decide how to assess the students in choosing assessment methods (AFT et al., 1990). Yet, the instructors have the knowledge of how to assess their students, but their knowledge is not technical, but practical. Though some instructors find pre-service assessment training effective, the instructors develop their assessment knowledge mainly by assessing and evaluating their students. They benefit from the practical knowledge in deciding how to assess. It is most probably because the instructors are aware of the inconsistency between ideal and real assessment practices and prefer to act depending on reality. Besides, the teacher should know the effects of different

assessment methods on decision-making (AFT et al., 1990). Similarly, the instructors are familiar with such effects since they know the strengths and weaknesses of assessment methods.

MAC (2013) also emphasized that the teacher should consider several factors to determine the appropriateness of assessment methods to their students. Similarly, the instructors give importance to their students by trying to choose an assessment method that the students are familiar with and capable of doing. Despite this finding, they consider practicality more than their students probably because of their concerns about workload.

Finally, the teacher should ensure that the factors like measurement error and validity do not influence his assessment results (MAC, 2013). Yet, the instructors do not consider measurement error and validity more in choosing assessment methods because of their concerns about practicality though they are familiar with validity and measurement error.

5.2.2. Developing appropriate assessments for instructional purposes

A teacher should be critical in assessment and evaluation practices, which is highly emphasized in the literature (e.g., Montee et al., 2013; Nier et al., 2013; Reistenberg et al., 2010; Walters, 2010;). In line with this suggestion, the Turkish instructors in this study adopt a critical attitude when they develop their exams. The instructors form their own criteria through which they assess, evaluate and choose vocabulary, reading passages and listening audio. They self-assess their questions in terms of use of language, wording, meaning and clarity after writing the questions. If the instructors want to select questions among the available ones including their course books and the Internet materials, they assess and evaluate such questions critically before using them on their exams. This critical attitude is utilized for the previous exams prepared by the instructors to determine if the previous exams can be re-used.

In addition to this finding, the instructors consider content validity more than other types of validity because their course books determine each step of developing assessments. This result is in parallel with the fact that what a teacher teaches and what a teacher assesses are consistent with one another (Brown, 2004; Yıldırım, 2012).

Besides, it is essential for the teacher to know his students' levels and do their assessments accordingly (Izci & Siegel, 2014; Munoz et al., 2012; Sezer, 2012). Corroborating this finding, the instructors always consider their students' levels of English in choosing vocabulary, reading passages and listening audio, writing their questions, assessing and evaluating the available questions and self-assessing the exams.

When developing their exams, the instructors use the assessment knowledge base that they acquire through experience. The assessment knowledge base influences the ways that the instructors make exams valid, develop their exams and evaluate the quality of their exams. This finding is confirmed by Gottheiner and Siegel (2012) who revealed that personal knowledge of assessment affects decision-making in evaluation. The instructors also develop personal assessment beliefs and goals based on their knowledge of assessment as found by Davison (2004), Izci and Siegel (2014) and Munoz and her colleagues (2012). For example, some instructors do not follow the mutually taken decision on using selected response as the method of assessment. Instead, they decide to use constructed response and performance assessment because these assessment methods are relevant to the instructors' personal goals and beliefs. Another example is that some instructors think listening is more difficult than reading for their students, so they prepare listening questions with three options and reading questions with four options. Therefore, it can be thought that experience is the basis of the instructors' assessment and evaluation practices as reported by Scarino (2013), Shohamy and her colleagues (2008) and Yıldırım (2012).

As mentioned, the instructors consider content validity more than the other types of validity. They develop several personal ways such as relating exams to course goals, using different types of questions and following assessment criteria strictly to make exams valid. Though Köksal (2004) and Sarıçoban (2011) found that content validity is a problem in the Turkish EFL teachers' exams, this result shows that the instructors in this study sought to make their exams content valid.

Another issue is the types of the questions used by the instructors. The instructors use multiple-choice, true-false, matching, fill-in-the-blank and close test in

their selected response exams. Consistent with this result, Öz (2014) revealed that such questions are the most preferred and used questions by the Turkish EFL teachers.

Finally, most of the instructors use teacher-made assessment materials to assess and evaluate their students. Being in parallel to the suggestion by AFT and its partner organizations (1990), this result indicates that they know the principles used for determining how to use and develop assessment in the class. Yet, the instructors do not provide any information about the quality of their exams in terms of validity and reliability when they develop the exams as opposed to the recommendations of JCTP (2002) and NCME (1995). Some instructors do not pay attention to the copyrighted materials in developing exams in contrast to the suggestion of NCME (1995).

5.2.3. Administering exams, scoring them and interpreting exam results

The Turkish EFL instructors follow a standard procedure in exam administration because they are observed to do the same things when they administer their exams. It indicates that the instructors understand and know the established procedures for administering an exam. Therefore, this finding is consistent with JCTP's suggestion (2002). The instructors inform their students about the exam rules, exams' structures, how to answer questions and when the exams will start and finish. They also allow their students to ask questions about the exams at the beginning of and during the exam. What the instructors do is in line with what JCTP (2002) and MAC (2013) recommended to apply assessments properly. In addition, JCTP (2002), MAC (2013) and NCME (1995) emphasized that a teacher has to provide the security of their exams and avoid any issue that may invalidate assessment results and misrepresent his students' real levels. Accordingly, the instructors are familiar with reliability and know its effects on assessment, so they take different precautions like making students turn off their cell phones to prevent cheating in exam administration and apply several ways like making exams content and face valid and double-checking exams to make their exams reliable. The instructors also consider confidentiality as important and seek to make their exams and exams' results confidential in several ways like preparing and printing their exams out of class time. This result corroborates the view that confidentiality is an essential part of assessment and evaluation (NCME, 1995). All of

the findings conform to the result that the Turkish EFL teachers' exams were improved in terms of reliability from 2004 to 2011 (Sarıçoban, 2011).

Scoring shows the weights given to each section and their items by a teacher on the exams (Brown, 2004). In parallel with this information, the instructors determine the weights given to listening, grammar, reading and vocabulary in ESP and/or EAP courses and score their exams in different ways to reflect these weights. Besides, they grade their selected response exams attentively because they do everything in grading their students' exams twice. If they ask open-ended questions, they prepare certain grading criteria and follow the criteria strictly. The instructors consider fairness and objectivity in grading selected response and constructed response exams. Therefore, they ensure the confidentiality of their grading and seek to lessen the effects of any factor irrelevant to the purposes of assessment on grading as recommended several studies in the literature (e.g., JCTP, 2002; MAC, 2013; NCME, 1995).

Providing consistent assessment results is as important as administering, scoring and grading. Therefore, the instructors apply different ways to obtain consistent assessment results. For instance, observing students can provide a basis to make decisions on their learning progresses (Davison & Leung, 2008; Herrera & Macias, 2015; Rea-Dickins, 2004). In line with this finding, some instructors in this study compare their students' grades with their observations to render their assessment results consistent. The instructors also consider fairness, objectivity, validity and reliability in their assessment and evaluation make assessment results consistent.

Considering validity, reliability, consistent assessment results, fairness and objectivity, most of the instructors pay attention to the washback effects of their exams even though they are not familiar with the term. They seek to create positive washback effects on their students through some ways including making their exams content and face valid and aligning their exams with their course objectives. This finding agrees with the explanation of Boyd (2015), Malone (2011) and Rogier (2014). Yet, it conflicts with Köksal (2004) and Sarıçoban (2011) who revealed that washback was a problem for the Turkish EFL teachers.

Most of the instructors interpret their assessment data in two ways. First, they use their assessment data to improve their students' learning by finding out their

strengths and weaknesses and helping the students to overcome weaknesses. Second, they assess their instruction, goals, objectives and exams so that they can improve their teaching. This result is confirmed by several studies (i.e., Chan, 2008; Herrera & Macias, 2015; Munoz et al., 2012; Rogers et al., 2007) in the literature. On the other hand, the instructors' personal beliefs about certain assessment methods prevent them from interpreting their assessment data as reported by Davison (2004) and Munoz and her colleagues (2012). In addition to this result, external factors including workload and the number of the students lead to the same problem as stated in the literature (e.g. Alkharusi et al., 2012; Ataman & Kabapınar, 2012; Büyükkarcı, 2014; Izci & Siegel, 2014; Özer & Karaoğlu, 2017).

As suggested by JCTP (2012), MAC (2013) and NCME (1995), a fair and reasonable procedure should be set up for students to make exam complaints. In line with this suggestion, some instructors react to their students' demands to make exam complaints depending on the type of the exams. However, some participants are affected by their students' attitudes and the number of the students who make exam complaints.

5.2.4. Using assessment results in making decisions about students, planning instruction and developing curriculum

A teacher should accumulate assessment data to make instructional decisions in several levels. Accumulated assessment data are evaluated in two ways: making instructional plans to improve students' learning and self-assessing instruction and assessment to improve teaching (AFT et al., 1990; Chan, 2008; Herrera & Macias, 2015; MAC, 2013; Munoz et al., 2012; Rogers et al., 2007). Just as stated in the literature, most of the Turkish EFL instructors in this study find out their students' weaknesses and seek to help the students to overcome the weaknesses when they evaluate their assessment data. Consistent with the finding told in the literature, the instructors also self-assess their teaching, goals, objectives and assessments to improve their teaching by finding out and overcoming their weaknesses. The instructors believe that they interpret their assessment data correctly considering their beliefs that (a) the teacher should be open to self-assessment and to make change, (b) the teacher and his students should be evaluated depending on the students' grades and real performances,

(c) the exams are valid and (d) assessment data are not always true. Several studies in the literature (e.g., Alkharusi et al., 2012; Ataman & Kabapınar, 2012; Aydoğmuş & Çoşkun Keskin, 2012; Izci & Siegel, 2014; Rogers et al., 2014; Scarino, 2013) confirm this finding as the instructors' beliefs clearly influenced their assessment and evaluation practices in the study. Though some instructors try to use their assessment data to improve their instruction and their students' learning, the others cannot do so because of several external factors including workload, syllabi and the number of students which are highly cited in the literature (e.g., Ataman & Kabapınar, 2012; Aydoğmuş & Keskin, 2012; Büyükkarcı, 2014; Eğri, 2006; Kuran & Kanatlı, 2009; Özer & Karaoğlu, 2017). The instructors also cannot make changes in the curriculum because it is a mutual and centralized one and they are not allowed to make changes, but if their curriculum is personal, they can make changes. This result is in line with what Kiomrs and his colleagues (2011) and Riazi and Razavipour (2011) revealed about the negative effect of centralized curriculum on teaching.

5.2.5. Developing valid grading procedures which use students' assessments

Much research in the literature (e.g., Scarino, 2013; Shohamy et al., 2008; Yıldırım, 2012) revealed that teaching experience affects a teacher's assessment and evaluation practices. In line with the finding, the Turkish EFL instructors in this study develop their grading procedures in two meetings under the influence of negative experience caused by their students as found by Davison (2004) in the literature. Their negative experiences prevent the instructors from using performance assessment, while they are motivated to choose selected response as a result of their positive experiences with it. The instructors have two quizzes, one midterm and one final exam in their grading systems because they believe that quizzes, midterm and final exams help them to maintain their teaching and encourage their students to study. Additionally, the other reasons of using midterm and final exams are that other universities use both commonly to evaluate their students and that they want to vary their assessments. They also believe that quizzes are part of language learning and enable their students to find out and overcome their mistakes. Several studies in the literature (e.g. Davison, 2004; Izci & Siegel, 2014; Scarino, 2013) corroborated this finding because the instructors' beliefs are the reflections of their teaching philosophies. In addition to this finding, other

studies in the literature (e.g. Inbar-Lourie, 2008b; Munoz et al., 2012; Rogers et al., 2007; Sever & İflazoğlu Saban, 2013; Yazıcı & Sözbilir, 2014; Webb, 2002) revealed that lack of time, the number of students and expectations from the teacher affect the teacher's decision-making in assessment. In accordance with this finding, the instructors mutually decide to decrease the number of quizzes from two to one. The instructors determine the percentages of each component in the grading systems depending on their content coverage. The instructors' grading systems are built upon their students' assessments because students' assessments make a grading procedure valid (AFT et al., 1990). Most of the instructors believe that the grades given through their grading systems are valid, fair and rational because of content validity, fair and objective grading, aligning the exams with the course objectives, considering students' levels of English and informing students about the assessment criteria. The grades also reflect their preferences and judgments. These findings are in agreement with AFT and its partner organizations' suggestion (1990). The instructors use their assessment results for two purposes. As Herrera and Macias (2015) stated, the instructors first decide who passes and fails based on the grades in the study. Second, some instructors self-assess their instruction in the study as reported by Chan (2008), Munoz and her colleagues (2012) and Rogers and his colleagues (2007).

5.2.6. Communicating assessment results to students and administrators

The Turkish EFL instructors in this study inform their students about the limitation, meaning and implication of their assessment results as well as the intended interpretation of the assessment results as recommended in the literature (e.g. AFT et al., 1990; JCTP, 2002; MAC, 2013). NCME (1995) also stated that a teacher should avoid the misinterpretations and misuses of assessment results by informing students about the possible effects of assessment and indicating them how to interpret the assessment results. In parallel with this, the instructors mention the possible effects of their assessments and show them how to interpret the assessment data when they inform their students, so they hope to avoid misinterpreting and misusing their assessment data. The instructors are also aware of the significance of measurement error and seek to prevent measurement error in communicating their exam results and making decisions depending on the results as explained in the literature (AFT et al., 1990). Though the

instructors are concerned about their workload, most of them try to use their assessment results as feedback for improving their students' learning and their instruction. In consistent with this finding, ITC (2001) and MAC (2013) strongly emphasized the importance of giving timely, descriptive and actionable feedback to students to help students use feedback and enhance their learning.

5.2.7. Recognizing unethical, illegal and inappropriate assessments and uses of assessment information

Validity, grading and students have shaped the understanding of ethical and legal assessment and evaluation practices of the Turkish EFL instructors in this study. The instructors think that validity, grading and students show their legal and ethical responsibilities in assessment and the instructors are aware of the effects of such responsibilities on their teaching, which is suggested in the literature (AFT et al., 1990). The instructors also know the limits of their responsibilities in the study as recommended in the literature (AFT et al., 1990). Thus, they believe that knowing their responsibilities, the effects of responsibilities and the limits of their responsibilities can avoid the harmful effects of assessment and evaluation on their students. In addition, NCME (1995) pointed out that a teacher should pay attention to the confidentiality of his students and know their rights as test takers. To do so, the instructors try to make their exams and exam results confidential and inform their students in each stage of their assessment in the study.

5.3. Factors Affecting the Implementation of Language Assessment Literacy

During the research, it has been found out that the Turkish EFL instructors are affected by several factors including workload, the number of the students and time in the study. These factors have an effect on one, some, most, or all of the sub-components of language assessment literacy.

The first sub-component is choosing appropriate assessment method for instructional purposes. The experiences of the previous instructors related to using the different types of assessment methods in the previous years, the instructors' concerns about a possible increase in their workload, the number of their students, lack of time

and teaching different courses in a week have influenced their decision-making in choosing assessment methods. Besides, their purpose to check what the students learn and do not learn have affected them. The mutual decision on the assessment method have created a kind of obligation for the new instructors to use selected response though some instructors have not wanted. On the other hand, some instructors have not obeyed the mutual decision because of their personal beliefs about the assessment tools and expectations from their students. In addition, the name, goal and content of a course have directed one of the instructors in choosing assessment methods. Making a judgment about the components of the elective courses' syllabi has also influenced her. In terms of validity, content validity has affected what most instructors understand from validity in the study. Besides, self-assessment, their willingness to make changes, personal beliefs and workload have effects on what the instructors understand from valid and invalid assessment data.

The second sub-component is developing appropriate assessments for instructional purposes. Their course books are the key determiner of every step of developing an assessment including writing/developing exam questions, choosing listening audio and/or reading passages and using the types of the questions because they shape what the instructors teach in their classes directly. The similarity between the course books and the exams affect the instructors a lot because they use the types of the questions similar to the course book activities and audio and passages similar to the ones in the course books in terms of topics, length, the students' levels of English and the target words. The instructors' personal beliefs about the assessment methods, personal goals, the name of the courses and its content have caused the instructors to use the certain types of the questions in their assessment and evaluation. Content validity, peer feedback and the levels of the students have impacted the ways the instructors use to make their exams valid. Time, workload and the number of the students help the instructors to decide how to prepare their exams. The weights given to listening, reading, vocabulary and grammar have affected the instructors' decisions about how many different sections will be in their exams and how many questions they will ask in each section. Their personal decisions are effective in the ways the instructors write their options for different sections, decide what to ask from listening audio and/or reading passage and choose words from their course books to ask. In

addition, their students' attitudes toward some parts of their courses help some instructors to decide what to ask in their exams. Brainstorming, outlining, private speech and code-switching have effects on their writing questions. Self-assessment is also effective in evaluating their questions and choosing listening audio and/or passages for their exams. Content coverage, understandability for the students and reliance on the course books have influenced how the instructors select questions among the available ones. In addition, the students' previous and future comments have an effect on how some instructors decide what to ask. Some of the instructors' pre-service assessment training have directed them in writing options for multiple-choice questions and preparing vocabulary questions. Besides, the types of cheating and the degree of insistence on cheating are the determiners of how the instructors react to the cheating students.

The third sub-component is administering exams, scoring them and interpreting their results. What the instructors understand from validity influences what they understand from reliability a lot in the study. Reliability, standardizing the administration of the exams, timing, informing students, not being distractive during the exams, the number of the students and avoiding cheating are effective in how the instructors administer their exams. The mutual decision about the number of the questions and the points given to each question, the distribution of ESP and EAP in three-hour academic English classes, the number of the questions in the exams and the difficulty levels of the questions determined based on the types of the questions are the determiners of how the instructors score their exams. The way how the instructors grade their exams is under the effect of the assessment methods chosen, double-checking everything, confidentiality and objectivity. Besides, the students' classroom performances, assessment criteria, fairness and objectivity in grading, content validity, relevance in scoring, cheating and measurement error have effects on the decisions given about interpreting assessment results consistently. The purposes of assessment are effective in interpreting the instructors' formal assessment results, while personal beliefs are effective in interpreting their informal assessment results. The instructors' willingness to do something based on their assessment results has a direct impact on the instructors in terms of using their assessment results to (a) improve their assessment tools, (b) help their students find out and overcome their weaknesses and (c) make their

exams have positive washback effects on their students. Self-assessment is also effective in the instructors' using assessment result to improve assessment tools as reported by AFT and its partner organizations (1990) and MAC (2013). The students' willingness to learn, time, their workload, selected response and the number of the students have influenced the instructors' use of assessment results to improve their students. Being aware of washback has an effect on making the instructors' exams have positive washback effects as mentioned by Rogier (2014). Besides, the types of the exams, the students' attitudes toward the instructors and the number of the students shape the instructors' attitudes toward the students' exam complaints.

The fourth sub-component is using assessment results in making decisions about students, planning instruction and developing curriculum. The instructors' evaluations of the students' grades are under the influences of self-assessing their instruction in line with AFT and its partner organizations (1990) and MAC (2013), the purpose of assessment and willingness to help their students. The correct interpretations of the assessment data are affected by the instructors' self-assessment as cited in the literature (e.g., AFT et al., 1990; MAC, 2013), willingness to change their teaching way, validity, following assessment criteria strictly, sharing and discussing the data together. In addition, workload, lack of time, syllabi, self-assessment, ideas about selected response and the difference between realism and idealism are influential in the instructors' decisions about using assessment results to develop instructional plans for their students. Besides, whether the curriculum is mutual or personal has affected whether the instructors can make changes in it because the instructors cannot make any change in the mutual curriculum, but can change their own curriculums based on their assessment results. Self-assessing instruction based on the students' grades helps the instructors to make changes in their teaching in the study. In addition, the number of the students, the teaching environment and workload have affected the instruction in decision-making about planning instruction.

The fifth one is developing valid grading procedures using students' assessments. Standardization, the experiences of the previous instructors, concerns about workload, legal obligation (for using midterm and final exams), (midterm and final exams') being generally used, personal beliefs about assessment methods, taking personal initiative based on experience, the name of the lesson, its content and the

classroom activities have influenced how the instructors have chosen the components of a grading system. The percentages of each component in the mutual grading system are determined under the effect of their content coverage and their purposes of use by the instructors, while the personal grading system is developed according to the weights of the classroom activities. In addition, personal judgments about the components of the grading systems designed by someone else lead to some changes in them. What the instructors think about the validity of the grades given through grading systems is under the influence of content validity, fairness and objectivity in grading, the levels of the students, selected response, explaining assessment criteria a lot and using the same exams at different times. Besides, legal obligation and willingness to self-assess have impacts on the instructors' purposes for giving grades.

The sixth one is communicating assessment results to students and administrators. The instructors' choices of communicating their assessment results are determined under the effect of their personal preferences, technology, confidentiality and legal obligation. Self-assessment, self-evaluation and evaluating someone are effective in deciding what the instructors' assessment results mean to their students and administrators. The fact that the interpretations are made based only on assessment results shapes the instructors' ideas about the correct interpretations of assessment results a lot. The standardization of the grading system, assessment criteria, selected response and the students lead to some limitations in interpreting the instructors' students' assessment data, which demonstrates that the instructors are aware of the limitations in interpreting assessment results in the study as stated in the literature (e.g., ITC, 2001; JCTP, 2002). Besides, the reason for failure and selected response influence the reflections of the instructors' interpretations of the assessment data. The students' misunderstanding of assessment criteria and the problems in teaching help the instructors to find some solutions to avoid misinterpretations. Besides, the validity and reliability of the exams influence the ways the instructors use to deal with measurement errors.

The seventh one is recognizing unethical, illegal and inappropriate assessments and uses of assessment information. Grading, validity, assessment criteria, students, reliability, the announcement of assessment results and confidentiality shape the

instructors' understanding of ethical, legal and appropriate assessments and uses of assessment information.

To sum up, the findings support the finding that teachers' knowledge of assessment tools and of assessment interpretation and action taking affect assessment practices (Gottheiner & Siegel, 2012). Besides, the findings indicate that the instructors' purposes of using assessment data are effective in assessment practices in the study, which is highly stated in the literature (e.g., Chan, 2008; Herrera & Macias, 2015; Malone, 2013; Mertler, 2003; Munoz et al., 2012; Rea-Dickins, 2004, 2006; Rogers et al., 2007; Saad et al., 2013; White, 2009). In addition, the findings reveal that the instructors' assessment values and beliefs influence their assessment practices, which is consistent with Davison (2004), Izci and Siegel (2014) and Scarino (2013). Besides, the instructors' previous experiences are also found to influence their implementation of language assessment literacy in the class as stated in Scarino (2013). In addition to these results, the findings demonstrate that the external factors including time, teaching materials (course books), the course objectives, the content of the course, workload, the number of the students, the levels of the students, curriculum and syllabi have certain effects on the instructors' implementation of one, some, most, or all sub-components of language assessment literacy in the class in the study. These findings corroborate with the findings of different studies which demonstrate that such external factors caused the participant teachers to have negative beliefs about the certain types of assessment and not to use such types in assessing and evaluating their students in the literature (e.g., Ataman & Kabapınar, 2012; Aydoğmuş & Çoşkun Keskin, 2012; Chan, 2008; Izci & Siegel, 2014; Jannati, 2015; Kuran & Kanatlı, 2009; Munoz et al., 2012; Özer & Karaoğlu, 2017; Rogers et al, 2007; Sezer, 2012). Moreover, consistent with Scarino (2013), the interaction between the instructors is effective in implementing language assessment literacy in the class.

As understood from the previous paragraphs, several factors have direct effects on how the instructors implement the sub-components of language assessment literacy in the English class. It is also important to know education and teaching approach (Abell & Siegel, 2011; Gottheiner & Siegel, 2012; Izci & Siegel, 2014; Siegel & Wissehr, 2011), assessment and evaluation approach (Gottheiner & Siegel, 2012; Inbar-Lourie, 2008a; Izci & Siegel, 2014; Leung & Lewkowicz, 2006) and assessment and

evaluation training (DeLuca et al., 2013; DeLuca & Klinger, 2010; Leung & Lewkowicz, 2006; Lomax, 1996; McGee, & Colby, 2014; Richardson et al., 2015) so that how some other factors become effective in the implementation of language assessment literacy in the class can be understood. First of all, the instructors' understanding of education and teaching aims to prepare the students for the future in the study. This affects what the instructors teach and do not to teach to the students, how they write options, why they want to assess the students and what they want to assess as mentioned commonly in the literature (e.g., Abell & Siegel, 2011; Gottheiner & Siegel, 2012; Izci & Siegel, 2014; Siegel & Wissehr, 2011). Secondly, the instructors use an eclectic teaching method very sensitive to the student-related factors in the study. The instructors reflect this sensitivity in the seven sub-components of language assessment literacy by considering their students while implementing the assessment literacy in the class. Third, the instructors' understanding of assessment and evaluation is related to checking their students' learning and evaluating their teaching. The instructors' understanding influences the choices of assessment methods, purposes of assessing the students and themselves and interpretations of the assessment results as highly stated in the literature (e.g., Gottheiner & Siegel, 2012; Inbar-Lourie, 2008a; Izci & Siegel, 2014; Leung & Lewkowicz, 2006). Fourth, the feelings which assessment and evaluation create for the instructors affect how they choose assessment methods, develop assessments, grade exams, avoid possible measurement errors and have their own understanding of legal, professional and ethical assessment behaviors in the study. Similarly, the associations which assessment and evaluation create influence the instructors' purposes of assessing their students and uses of assessment and evaluation in their teaching. The instructors' previous experiences with different assessment methods when they were students cause the instructors to form certain ideas about and attitudes toward certain assessment methods, which leads to choosing certain assessment methods which some instructors believe are the correct assessment tools in their assessment systems in the study. This finding agrees with Izci and Siegel (2014) and Stiggins (1995). Besides, four instructors have pre-service training in assessment and evaluation, but three of them think it is effective because the instructors have learnt how to prepare exams, to make the exams valid and reliable, to judge the quality of the exams and to adjust the difficulty levels as several studies in the literature revealed (e.g.,

DeLuca et al., 2013; DeLuca & Klinger, 2010; Karaman & Şahin, 2014; Lomax, 1996; McGee, & Colby, 2014; Richardson et al., 2015; Yetkin, 2015). Yet, the last instructor does not think it is effective because he thinks it is too theoretical and lacks practice, which corroborates much research in the literature (e.g., Hatipoğlu, 2010, 2015a; Hatipoğlu & Erçetin, 2016; Popham, 2004; Stiggins 1991, 1995). The other instructors in the study do not have any pre-service training in line with the results of several studies in the literature (e.g., Hasselgreen et al., 2004; Montee et al., 2013; Vogt et al., 2008), but pedagogical formation training which the instructors find ineffective since the course teachers' attitudes toward the course were negative and pedagogical formation training was too theoretical and lacked practice, which is commonly cited in the literature (e.g., Hatipoğlu, 2010, 2015a; Hatipoğlu & Erçetin, 2016; Lam, 2015; Lomax, 1996; Kahl, Hofman, & Bryant, 2013; Koh & Velayutham, n.d.; Popham, 2006; Stiggins, 1991, 1995; Webb, 2002). Only one of the instructors in the study finds in-service training effective as he has learnt how different assessment methods work, but the other instructors having in-service training do not find it effective, which conflicts with the claim of Montee and her colleagues (2013), Nier and her colleagues (2013), Riestenberg and her colleagues (2010) and Walter (2010) that in-service training affects teachers' language assessment literacy positively, but agrees with the fact that in-service training does not affect assessment beliefs and practices (Büyükkarcı, 2014). However, the instructors' real improvement in assessment in the study is closely related to self-improvement through peer feedback (Munoz et al., 2012; Scarino, 2013; Tahmasbi, 2014), peer observation (Scarino, 2013; Tahmasbi, 2014), self-assessment (AFT et al., 1990), integrating theory with practice, using their previous assessment experiences (when they were students) (Izci & Siegel, 2014; Stiggins, 1995), research, self-interest and studying for CELTA and/or KPSS. The most important way to self-improvement for the instructors in the study is gaining experience because experience affects teachers' assessment beliefs and practices, which enables teachers to learn how to assess and evaluate by assessing and evaluating the students (Alkharusi, 2011a, 2011c, 2011d; Chan, 2008; Eğri, 2006; Hasselgreen et al., 2004; Hatipoğlu, 2015b; Guerin, 2010; Mede & Atay, 2017; Vogt et al., 2008; Witte, 2010). As a result of self-improvement, the instructors in the study have become more autonomous and independent in deciding what to do in assessment and evaluation, creative in writing questions, teacher-centered

in grading, student-centered in preparing questions, specific in deciding what to ask, flexible in evaluating the students' performances and open to using different assessment methods.

5.4. The Effects of Language Assessment Literacy on the Turkish EFL Instructors

Teachers use assessment data to assess their instruction and to check and improve their students' learning (AFT et al., 1990; Chan, 2008; Munoz et al., 2012; Rogers et al., 2007; White, 2009). Therefore, valid and invalid assessment data cause most instructors in the study to self-assess their instruction, goals and course objectives to find out and overcome the problems in their teaching. As a result of self-assessment, the instructors evaluate their courses by checking whether they achieve what they have planned and decide whether they will go on using the same teaching methods or change the methods as reported by AFT and its partner organizations (1990). The instructors in the study also use valid/invalid assessment data to help the students to find out and overcome their weaknesses. In addition, the standardization of the grading system creates an obligation for the instructors in the study to use it without making any change, so they have to follow the grading system strictly though some instructors make several changes in it. According to most instructors, this whole process makes them believe that their interpretation of assessment data is consistent in the study, which is in parallel to the finding of Davison (2004) and Scarino (2013). Besides, the instructors are concerned about the grades because they want to be fair and objective in grading the exams. The instructors' concerns cause them to choose selected response which shows only what their students know and do not know and which some instructors believe makes the grades given through their grading system valid in the study. Therefore, the instructors are relieved as they believe selected response reduces the number of the students' complaints about their exams and reduce their workload as mentioned by Sezer (2012).

The previous assessment experiences the instructors gained when they were students cause the instructors to form negative and positive attitudes toward certain assessment methods, which is confirmed by several findings in the literature (e.g., Davison, 2004; Hatipoğlu, 2015a; Izci & Siegel, 2014; Stiggins, 1995). In addition, the instructors have gained experiences in assessment and evaluation since they started to

work as teachers. According to some instructors, their previous and new assessment and evaluation experiences give them right to take initiative in their assessment and evaluation and to select some other methods in order to assess the students. Their experiences also help these instructors to determine the components of their grading systems in the study. Meanwhile, pre-service assessment training and previous and new assessment experiences create the instructors' assessment knowledge which is concurred with Abell and Siegel (2011) and Gottheiner and Siegel (2012). Most instructors in the study use this knowledge of assessment to evaluate the available questions for preparing exams. The instructors reflect their purposes by taking initiative in choosing assessment methods in the study. The experiences also affect the instructors to form negative attitudes toward the assessment methods which they favor and positive attitudes toward the assessment methods which they do not favor. The instructors who choose different assessment methods ignore their concerns about grading and follow their own decisions strictly. The instructors interpret their assessment data depending on their personal goals. The instructors either help their students to find out and overcome their weaknesses or self-assess their teaching by interpreting assessment results, which supports that assessment data should be used to improve instruction and students' learning as several studies in the literature pointed out (Chan, 2008; Munoz et al., 2012; Rogers et al., 2007). Some instructors do both by interpreting assessment results. The instructors interpret formal and informal assessment results depending on their personal goals. Besides, their personal goals help the instructors to determine the components of the grading system and partly the percentages of the components in the total grade. Their goals help the instructors to explain the validity of the grades given through their grading systems to their students and administrators. In addition to their personal goals, their personal beliefs influence what the instructors will and will not teach in the class and what they will and will not ask in the exams as indicated by Davison (2004) and Scarino (2013). The instructors in the study choose the types of the questions to use in the exams depending on their beliefs. The instructors make their exams valid and reliable by using some ways like using different types of questions in an exam under the effects of their personal beliefs. Some instructors also decide the number of the options in the listening and reading questions through the belief that their students find listening more difficult than reading in the study. In addition, their personal beliefs affect how the

instructors evaluate assessment data in the study. According to some instructors, following the assessment criteria strictly provides them with the correct evaluation of assessment results. Depending on their preferences, some instructors take initiative in their quizzes and use constructed response as well as selected response. The instructors also decide how to prepare questions with the help of their preferences. They deal with their students' exam complaints and make changes in their instruction depending on their preferences. They announce their grades to their students by using the ways they prefer.

The instructors' assessment and evaluation are course book-centered in the study. Their course books determine the types of the questions to prepare in the exams, the choice of the assessment methods, the selection of the components of the grading system (together with the legal obligation to use one midterm and final), the content of the exams, the development of the questions in the exams, the selection of audio and passages to use in the exams and the number of the questions in the different sections of the exams. To decide the number of the questions, the instructors use the course books to determine the weights given to different sections in the study as Brown (2004) suggested in how to decide a scoring plan for an exam. Consequently, the instructors reflect these weights by asking questions in a different number in different sections, by giving each question the same point, or by scoring the questions differently. In addition, some instructors determine the percentages of the components in their grading systems by the weights given to the classroom activities in the course books. Meanwhile, most instructors decide the percentages of the components in the grading systems by the content coverage of each exam they determine through the course books as mentioned by Brown (2004). Most instructors believe that this way makes the grades given via their grading systems valid in the study. As a result of being course book-centered, the ways the instructors use to make their exams valid and reliable are shaped by the course books. Accordingly, the instructors use the types of the questions similar to the ones in the course books. They believe this reduces their workload, avoids their students' anxiety, makes their exams face valid and affects scoring the questions in their exams. Besides, the instructors decide how to start their exams, brainstorm and outline their exams, determine what to ask and write the content of their questions through their course books. They also evaluate the available questions by using their course books. In

addition, most instructors develop their questions by adapting the exercises in the course books and/or prepare their questions by taking and using the exercises in the course books without making any change like the participants did in the Kiomrs and his colleagues' study (2011).

In addition to being course book-centered, the instructors' assessment and evaluation practices are student-centered because they want to have some data to evaluate their students' learning in the study. According to Chan (2008), Munoz and her colleagues (2012), Rogers and his colleagues (2007), such assessment data provide teachers with the concrete results of the students' learning to evaluate and to base their evaluation on. The instructors in the study first pay attention to their students' levels of English. They choose their assessment methods which they believe their students can do. They select and use the types of the questions which their students are familiar with and the exam listening audio and reading passages which are similar to the ones used in their course books in terms of topic, length and target words. They take into account their students' levels of English in making decisions about choosing and using the types of the questions and the exam listening audio and reading passages. They also try to find the audio and passages which their students can understand. They write their own questions, self-assess the questions after writing them and/or evaluate the available ones depending on the levels of the students. In addition, they decide what to ask by referring back to what their students have learned in the class. As their assessment and evaluation are student-centered, they inform their students about their assessment criteria, grading system, courses, exams and exam rules. When they prepare their own questions, they pay attention to their students' previous comments about their exams. They administer their exams by informing their students about the exam, exam rules and duration, helping them if the students have any question and not disturbing the students during the exam, which shows they give importance to their students. As a consequence of being student-centered, the ways the instructors make their exams valid and reliable are also determined by the factors related to their students. According to some instructors, being student-centered improves the validity of the grades given through their systems. Besides, their students' attitudes toward and comments about the exams, together with the types of the exams, determine the ways the students of some instructors follow in order to make complaints about their exams results.

Most instructors have improved their assessment knowledge by interacting with and observing their colleagues, which is in line with several studies in the literature (e.g., Munoz et al., 2012; Scarino, 2013; Tahmasbi, 2014). This causes the instructors to give importance to peer feedback in which the instructors share their exams with each other, answer the questions, check the wording and the use of language in the questions and control the answer keys, so the instructors can give and receive feedback to make some necessary changes in their exams, which they believe makes the exams valid and reliable in the study. Like peer interaction, self-improvement and pre-service assessment knowledge are effective in the instructors' assessment practices as suggested by much research in the literature (e.g., DeLuca et al., 2013; DeLuca & Klinger, 2010; Karaman & Şahin, 2014; Lomax, 1996; McGee, & Colby, 2014; Richardson et al., 2015). Self-assessment and pre-service training are used as determiners in writing the options for multiple-choice questions in terms of length and degree of challenge and in using the same parts of speech in vocabulary question in the study.

The instructors decide how to prepare exams (writing or selecting questions) because they may or may not have enough time to prepare questions for the exams. Time also makes the instructors reduce the number of the sections in the exams. It is one of the key determiners used to choose the exam listening audio and reading passages by creating a similarity to the ones used in the course books in terms of duration and length. The instructors give importance to time to make their exams reliable by giving their students enough time to study for the exams and to finish the exams, which helps their exams to have a positive washback effect on the students. In addition, the instructors score their questions depending on the number of the questions in the exams to make grading the exams easier as a result of the importance given to time.

The instructors consider self-assessment as vital in assessment and evaluation. They self-assess their instruction, exams, goals and objectives by using assessment data because assessment data enables teachers to have concrete results of their instruction, so teachers can check whether they achieve their goals, which is highly reported in the literature (e.g., AFT et al., 1990; Chan, 2008; Herrera & Macias, 2015; Munoz et al., 2012; Rogers et al., 2007; White, 2009). Self-assessment enables the instructors to choose exam listening audio and/or reading passages, to decide the quality of the

questions they write on their own, to select questions among the available ones, to determine what and how to ask and to find out and overcome their weaknesses. Thus, they can improve their teaching and assessment tools. The instructors believe that if they self-assess their teaching in terms of the ways aforementioned, they can interpret assessment results correctly. Therefore, most instructors can make instructional plans to improve their students.

Besides, validity, especially content validity has several effects on the instructors. The ways most instructors use to make their exams valid are also the ways used to make their exams reliable. Therefore, validity directly shapes how the instructors prepare their exams in the study. According to some instructors, it also improves the validity of the grades given through the grading system. In addition, most instructors believe in the study that it helps them to interpret assessment results correctly. In addition, reliability affects especially how the instructors prepare, administer and grade the exams. They prepare questions and answer keys carefully. They also do everything in grading the exam papers twice, so they believe that their grading is objective and fair. Through these ways, they think they avoid possible measurement errors in the exams. Therefore, they consider their interpretation of their assessment data as consistent. The instructors try not to disturb the students during the exams. They inform their students about their grading system, assessment criteria, courses and courses' goals, which some instructors think improves the validity of the grades given through the grading system. Most instructors pay attention to writing instructions in one, some, most, or all of their exams and print out the exams clearly so that their students can easily read the questions and do not have any problem because of the photocopied exams, all of which make their exams face valid. They take several precautions to avoid their students' cheating in the exams as suggested by MAC (2013). The ways the instructors use to provide the reliability and validity of their exams show how measurement errors affect the instructors' assessment and evaluation practices.

The instructors give importance to the consistency of their interpretations of assessment results. This importance makes some instructors compare their students' grades with their classroom performances which they determine based on observing the students in the class. It also makes some instructors use the same type of assessment method during the whole term.

Grades make most instructors interpret the students as successful and unsuccessful because they determine who passes and fail based on the students' grades owing to the legal obligation as stated by Herrera and Macias (2015). The grades show the instructors what their students have learned, but not whether the students can use what they have learned. Some instructors evaluate their courses by self-assessment depending on their students' grades as suggested in the literature (AFT et al., 1990), yet they know evaluating themselves based only on grades is not correct. Some instructors use assessment data to help the students to find out and overcome their weaknesses. In addition, the instructors use grades to evaluate their course books to decide whether they will go on using the course books or change them with the new ones. Similarly, some instructors use grades to improve their assessment tools by self-assessing their instruction and to check whether their instruction has become successful in teaching what they have planned to teach. This is in line with the suggestion and finding of several studies in the literature (e.g., AFT et al., 1990; Chan, 2008; Herrera & Macias, 2015; Munoz et al., 2012; Rogers et al., 2007; White, 2009).

The instructors' assessment and evaluation practices reflect their education and teaching practices in the study, which is confirmed by much research in the literature (e.g., Kiomrs et al., 2011; Leung & Lewkowicz, 2006; Leaph et al., 2015; Riazi & Razavipour, 2011). The instructors accept in the study that they are product-oriented because of the numbers of the students, workload, students' levels of English and lack of time although they want to follow their students' learning processes. In addition, the instructors educate and teach their students according to chosen assessment methods. Some instructors can reduce the number of the questions which they have found out their students are not good at through the students' previous exam results, so they increase the number of the questions which the students are good at.

Moreover, the instructors pay attention to the confidentiality of the exams and exam results as recommended by NCME (1995). How the instructors prepare the exams and announce the grades is shaped by confidentiality.

Most instructors are aware of the washback effects of the exams on their students, which is consistent with several studies in the literature (e.g., Boyd, 2015; Brown, 2004; Malone, 2011; Rogier, 2014). The instructors make the exams have

positive washback effects on the students through the ways used to make the exams valid and reliable.

What the instructors do during assessment and evaluation shapes their understanding of the ethical, legal and professional issues in assessment and evaluation. The ways to deal with the unethical, illegal and unprofessional issues are determined by their understanding of such issues.

5.5. Difficulties Encountered in Implementing Language Assessment Literacy and Ways to Overcome Them

The instructors in the study have difficulty in choosing assessment methods, deciding the number of the components in the grading systems, preparing exams, scoring them as they want and using the results in making decisions about the students and themselves due to their workloads. They have to choose selected response which is the least labor-intensive for them. Like choosing assessment methods, they also determine to use one midterm, one quiz and one final to reduce workload. Some instructors do not have enough time to write their own questions, so they have to choose questions among the available ones to save some time. Similarly, some instructors do not have enough time to develop instructional plans for the students by using assessment data. Consequently, they may ignore whether their assessment data are valid or invalid. In addition, workload prevents some instructors from interpreting assessment data to improve their assessment tools and to make changes in their instruction. Though some instructors want to check the exams with their students, workload does not allow them to do so. All findings indicate that the instructors in the study prioritize practicality and tend to apply less labor-intensive in their assessment practices. Some instructors give up their personal beliefs and goals to do so.

The high number of the students causes the instructors to experience some difficulties in choosing assessment methods, preparing and administering the exams and interpreting assessment results. The instructors have to choose the assessment method which is easy to grade and administer, so they choose selected response. The high number of the students is a big challenge for the instructors to administer exams, so they need to (a) prepare two types of booklets, (b) warn the students seriously about the

exam rules, (c) make the students turn off the cell phones and remove the notes on the desks, (d) arrange the students' seating order and (e) find someone else who helps them to proctor exams in order to avoid cheating as suggested by MAC (2013). Moreover, the high number causes some instructors not to be able to find out their students' weaknesses, so they cannot develop an instructional plan for the students. In addition, those instructors cannot interpret assessment results to improve their assessment tools and make changes in their instruction. Though some instructors check the exams with the students, the high number avoids the rest's doing so. As a result of the high number of the students in the class and workload, some instructors believe that their working environment is a barrier in interpreting assessment data to improve their assessment tools and to make changes in their instruction. These results reveal that the instructors in the study take strict precautions to avoid cheating and make practical decisions to overcome such problems. Some instructors prefer doing what they do not want to do to overcome the difficulties the high number of the students leads to.

The syllabi the instructors have to follow in the class require them to choose certain assessment methods in the study. The syllabi consume the instructors' time a lot as they need to cover the things on the syllabi in a certain time period. Therefore, some instructors cannot develop any instructional plans for the students and write their own questions to prepare the exams. In addition to the syllabi, the instructors use a mutual curriculum. The mutual use of the curriculum prevents most instructors from making decisions in improving the mutual curriculum by using assessment data even though some can make changes in the curriculum which they have prepared on their own. To overcome the difficulty which syllabi lead to, some instructors prefer assessing and evaluating their students without interpreting assessment data.

Lack of time causes difficulty in choosing assessment methods. The instructors cannot choose their favorite assessment methods as these methods are time-consuming and labor-intensive, so the instructors prefer selected response which they think saves their time in the study. Some instructors cannot use assessment results for improving their instruction and assessment tools and developing instructional plans for the students to find out and overcome their weaknesses owing to lack of time. Some participants cannot write original questions because they do not have enough time for it. All

findings point out that the instructors consider practicality superior to their personal beliefs and goals.

The explanations above indicate that these external factors made the instructors reluctant to use the certain types of assessment tools like performance assessment, use time-saving assessment methods and caused some instructors to use assessment data ineffectively. This finding corroborates with the findings of several studies in the literature (e.g., Alkharusi et al., (2012); Ataman & Kabapınar, 2012; Aydoğmuş & Çoşkun Keski, 2012; Büyükkarcı, 2014; Chan, 2008; Izci & Siegel, 2014; Kuran & Kanatlı, 2009; Özer & Karaoğlu, 2017).

Some instructors believe in the study that the standardization of the grading system in the academic English classes is an obligation to be fair and objective in grading. However, the standardization of the grading system is influenced by the experiences of the previous instructors in the department more as mentioned in the literature (e.g., Scarino, 2013; Shohamy et al., 2008; Yıldırım, 2012). The old instructors have the priority in making decisions about choosing assessment method. This restricts the new instructors as they want to change the grading system, but they cannot. The old instructors have a lot of negative experiences with performance assessment (presentation, report writing and evaluation grades) because of the students, which is in line with the finding of Davison (2004). In addition, the instructors are aware of the weaknesses of selected response as recommended in the literature (AFT et al., 1990), but they choose selected response to avoid the problems the old instructors experienced and to standardize the grading system. Standardizing the grading system also affects how some instructors score their exams. Some instructors want to score their exams according to the weights given to the different sections in their course books and to the difficulty levels of the questions, but they can not do so because the standardization process also determines how many questions will be asked in the midterm and final exams and which points will be given to each question. Besides, some instructors think the obligation of using the same grading system limits interpreting assessment results if the obligation is not followed strictly. The findings reveal that the instructors do not think for themselves, but follow the crowd in their assessment.

The assessment method the instructors have chosen is mainly selected response which causes some difficulties in interpreting and using the students' assessment data in making decisions about the students and themselves. The instructors have only information about what their students have and have not learned, so they cannot know whether their students can use what they have learned. Similarly, some instructors cannot use their assessment results to find out the students' weaknesses. As a result, this prevents some instructors from using assessment data to develop instructional plans for the students. Besides, some instructors believe selected response reduces the validity of the grades given through the grading systems, so they think it limits their interpretation of assessment results, which indicates that these instructors know the limitations of interpreting assessment data as recommended in the literature (ITC, 2001; JCTP, 2002). Consequently, some instructors do not believe that they can really check whether their instruction has become successful in teaching what they have planned. The findings reveal that concerns about workload and time force the instructors to choose and use selected response in their assessment and discourage some instructors from evaluating and using assessment results in a critical attitude.

The possibility of cheating is a threat to validate assessment data (MAC, 2013) because it causes teachers to evaluate themselves incorrectly based just on their students' grades even if they know high grades do not show that the students study a lot. Cheating also leads to difficulty in administering the instructors' exams. The instructors have to develop certain ways to deal with the cheating students without disturbing the other students during the exams as reported by MAC (2013).

The grades given by the instructors mean success and failure to their students. According to some instructors, their students blame the instructors for their failure. This finding is corroborated with the result of Hidri (2015) who found that teachers were blamed for the failure of an educational program. The instructors' students may want the instructors to change the grading systems if they fail. Depending on the grades, the instructors' students may have wrong expectations about themselves. Some instructors think that grades mean teacher and program evaluation based on the grades to their administrators, which is consistent with Popham (2011), but this limits the interpretation of their assessment results as reported by ITC (2001) and JCTP (2002). Therefore, it makes the evaluations based only on the grades wrong for some

instructors. The results make the instructors be ready to defend themselves against criticism on the evaluation made based only on grades.

5.6. The Implementability of Language Assessment Literacy

Most Turkish EFL instructors in the study define assessment literacy as the fact that a teacher knows what he does in assessing and evaluating his students as understood from the first interviews. They think that the teacher improves his assessment knowledge by training, reading an assessment source and assessing and evaluating his students. According to the instructors, the assessment knowledge acquired and learnt helps him to be familiar with assessment and evaluation, to monitor his assessment and evaluation processes, to evaluate his students through his own criteria, to organize the layouts of his exams, to integrate his knowledge with his teaching context, to understand his assessment results and to use technology in assessing and evaluating his students. Moreover, the last two instructors think that assessment literacy is that the teacher assesses and evaluates his students based on his course goals and objectives by being objective in grading, avoiding measurement errors and reflecting the weights given to different sections in the course book in the exams.

In addition to the instructors' definitions, Paterno (2001), Popham (2004) and Stiggins (1991) related assessment literacy to knowing what a sound assessment is. Knowing includes being familiar with the basic principles of sound assessment, using this knowledge in developing and using assessment methods and evaluating the quality of the assessments prepared (Paterno, 2001). Besides, Abell and Siegel (2011) thought assessment literacy is based on a teacher's view of learning and uses the interaction between the view of learning and the knowledge of assessment. It is also having the knowledge of how to assess what students have learned, to interpret assessment data and to use assessment data to improve students' learning and education program (Webb, 2002). The explanations reveal that the teacher should have the assessment-related knowledge, skills and competencies in order to be assessment-literate (Kahl et al., 2013). Similarly, Inbar-Lourie (2008a), Lam (2015), Malone (2013), O'Loughlin (2013), Pill and Harding (2013), Scarino (2013), Vogt and Tsagari (2014) and Tsagari and Vogt (2017) mentioned that language assessment literacy is having the knowledge of the basic principles of general assessment and evaluation, integrating it with the

knowledge of second language teaching and using this integration in choosing and developing assessment methods, interpreting assessment results and making decisions for improving students' learning and language teaching program.

The instructors' own definitions and the definitions of the scholars indicate that a language teacher should have the knowledge of assessment and evaluation and use the knowledge in the seven sub-components of language assessment literacy. As the findings and discussions show, each instructor in the study implements all sub-components of language assessment literacy in the English class. The instructors' understanding of assessment literacy reflects how they have improved their assessment and evaluate knowledge and how and why they use this knowledge in assessing and evaluating the students. Meanwhile, the instructors have some difficulties which prevent them from doing some of the requirements of the sub-components of language assessment literacy like not being able to use the assessment results to develop instructional plans for the students because of workload, the number of the students and lack of time. Despite these difficulties, the instructors in the study can implement their language assessment literacy in the class as the findings show.

5.7. Conclusion

This chapter has mentioned the discussion by relating the findings to the literature review based on the research questions. The next chapter is going to present the overview of the study, its main findings, its implications, its strengths and limitations and suggestions for further study.

CHAPTER SIX

6. CONCLUSION

6.1. Introduction

This part first presents the overview of the study by mentioning its methodology and main findings. It second explains its implications. Then, the part mentions its strength and limitations and ends up with the further research suggestions.

6.2. Summary of the Study

The present study provides detailed information about the implementation of language assessment literacy by the Turkish EFL instructors in the class in a specific teaching context. It also explains the effects of different factors on the implementation of language assessment literacy, the effects of language assessment literacy on the instructors and the difficulties encountered by the instructors. Consequently, the findings of the study allow to understand language assessment literacy in its epistemological, ontological and practical dimensions. The epistemological dimension explains how language assessment literacy ability is acquired and developed and how it turns into knowledge of assessment. The ontological dimension presents how language assessment literacy ability develops language teachers' educational being as the agent of language assessment. The practical one mentions how language teachers act upon the ability of language assessment literacy.

6.2.1. The epistemological dimension of language assessment literacy

Different definitions of language assessment literacy (e.g., Fulcher, 2012; Inbar-Lourie, 2008a; Lam, 2015; Malone, 2013; O'Loughlin, 2013; Pill & Harding, 2013; Scarino, 2013) perceive language assessment literacy as the ability to choose, develop, administer, score, use, analyze and interpret any language assessment. Consistent with various definitions in the literature, the Turkish EFL instructors in this study seem to

have the language assessment ability as they can implement the seven aforementioned components of language assessment literacy. However, the findings do not imply that these instructors are familiar with and know the social, historical, philosophical and political frameworks which are necessary for the literacy. Fulcher (2012) proposed that these four frameworks explain how different assessment practices affect teachers' assessment practices and are crucial to be more language-assessment-literate.

As revealed in the literature (Lam, 2015; O'Loughlin, 2013), teachers' ability to understand, acquire and master the skills, knowledge and principles of test-related issues like test construction, use, validity and reliability is a requirement for gaining language assessment literacy. Similarly, the EFL instructors in the study have the skills, knowledge and principles of test-related issues. They mostly have acquired and mastered these skills through gaining experience by assessing and evaluating their students. Yet, this finding conflicts with Inbar-Lourie (2008a) who objected to learning with trial and error method and emphasized the importance of training in gaining the ability of language assessment literacy. She claimed that the skills, knowledge and principles require knowing and being familiar with the modern theories of learning and assessment, language teaching pedagogy and theories of language.

Another point of the study is that the EFL instructors in the study mostly built their knowledge of language assessment upon their experiences in the process. When they started to work, they did not know much about language assessment and evaluation. Thus, they applied to different sources to develop their own way of language assessment. While some instructors were assessing and evaluating their students in the same way that the instructors were assessed when they were students, the others were doing peer assessment and integrating other colleagues into the process. Receiving some feedback, using observation technique and sometimes supporting their colleagues' assessment practices with their fully-tested ideas, the instructors help each other's development. Also from the findings, it is clear that having theoretical knowledge does not guaranty the best practices of assessment. Some instructors knew the theoretical aspects of language assessment, but lacked practice. They tried what they knew theoretically on the job and learnt how to use the theoretical knowledge in real teaching context. The second essential source is assessment beliefs for the instructors to develop assessment knowledge. Assessment beliefs caused some instructors to believe a certain

type of assessment worked better than others after they used different assessment methods. Assessment beliefs were so effective that these instructors made and followed their own assessment decisions despite the mutual decision on the chosen assessment method. This finding is concurred with Scarino (2013) who found out that teacher's beliefs are as effective as experiences in the development of language assessment literacy. In addition, the effects of beliefs and experiences are more obvious in the way the instructors define reliability, validity and measurement error than the effects of theoretical knowledge. The ELF instructors in the study developed basic and practical explanations for these basic principles of assessment as well as the ways they used to make the exams reliable, valid and free from measurement errors. All in all, being critical is an indispensable part of language assessment as supported by Lam (2015) and O'Loughlin (2013). Accordingly, it is very evident from these findings that the instructors mostly developed a critical attitude toward assessment based on experience and beliefs instead of theoretical knowledge.

Finally, why, what and how questions are significant to answer for doing a sound assessment, and the language assessment literacy ability helps to answer these questions. (Inbar-Lourie, 2008a). According to her, the why question expresses the rationale of assessment. To set up the rationale of the exams, the EFL instructors in the study mostly used assessment and evaluation for three purposes: deciding who passed and failed, checking the students' learning to enhance their learning and self-assessing their instruction to improve it as reported by Herrera and Macias (2015). The what questions describes and decides a trait to be assessed by knowing the modern theories of learning and assessment, language teaching pedagogy and theories of learning (Inbar-Lourie, 2008a). Yet, the EFL instructors in the study depended on their course books and classroom activities heavily and decided the traits to be assessed by setting aside this requirement. The how question indicates how appropriate assessments are developed to assess the trait and requires using the familiarity with the modern theories of learning and assessment, language teaching pedagogy and theories of learning (Inbar-Lourie, 2008a). However, the instructors developed appropriate assessments to assess the traits by using the knowledge of language assessment which they built upon experience and belief. In sum, it is certain from these findings that theoretical knowledge does not work as effectively as experience- and belief-based knowledge of

assessment as the second type of knowledge provides teachers with what they need in real teaching contexts.

6.2.2. The ontological dimension of language assessment literacy

The Turkish EFL instructors in this study know that language assessment creates personal responsibilities like developing assessments and keeping exams confidential as well as mutual responsibilities like choosing assessment methods and assessing their colleagues' exams. Such responsibilities constitute a big part of the instructors' teaching; therefore, their educational being. This finding indicates that the instructors were considered as the agents of language assessments as stated in Rea-Dickins (2004). Being the agent of language assessment is significant for teachers to be more language-assessment-literate because such a being keeps teachers responsible for every assessment-related activity, which is highly cited in much research in the literature (e.g., Alas & Liiv, 2014; Boyd, 2015; Davison & Leung, 2009; Newfields, 2006; Pill & Harding, 2013).

As explained in the literature (Herrera & Macias, 2015; Pill & Harding, 2013), agency in language assessment considers using assessment data as an essential necessity and part of language assessment literacy. This necessity strongly supports the use of assessment data to improve students' learning and teachers' instruction and opposes using data for giving grades. Consistent with this, most EFL instructors in the study used their assessment data to self-assess their instruction and assessments so that they could improve their instruction and assessments. These instructors also used assessment results to enhance their students' learning by helping the students to find out and overcome their weaknesses in the study.

Being the agent of language assessment also requires teachers to deal with the basic principles of language assessment such as validity and reliability. The EFL instructors' experience-based knowledge of assessment helped the instructors to develop and use ways for assessing the validity and reliability of their exams and avoiding any possible measurement error in the study. The experience-based knowledge of assessment has caused the instructors to be course book-centered, so the instructors do almost every step of language assessment depending on the course books. Even

though being not familiar with washback, most instructors try to create positive washback effects on their students through the experience-based knowledge of assessment. Therefore, the instructors' knowledge of assessment enables them to identify good and bad assessment as well as the positive and negative effects of their exams in the study as recommended by Boyd (2015). These results also indicate that the instructors in the study understand the importance of assessment in language teaching and that they try not to cause any validity and reliability problems in the study as suggested in the relevant research by several scholars (Alas & Liiv, 2014; in Shohamy et al., 2008). In conclusion, the findings make it clear that experience and beliefs can provide language teachers with enough knowledge to achieve the requirements of being the agent of language assessment literacy.

6.2.3. The practical dimension of language assessment literacy

The findings of the study indicate that the Turkish EFL instructors are very concerned about external factors including workload, the number of the students and syllabi. These factors are highly cited in the relevant literature by several researchers (e.g., Alkharusi et al., 2012; Ataman & Kabapınar, 2012; Aydoğmuş & Çoşkun Keskin, 2012; Büyükkarcı, 2014; Chan, 2008; Izci & Siegel, 2014; Kuran & Kanatlı, 2009; Özer & Karaoğlu, 2017). Therefore, the instructors prioritize practicality more in each sub-component of language assessment literacy in the study.

The instructors have chosen selected response as the official assessment method in academic English course in the study. They have experienced that selected response is time-saving and easy to administer and grade though they know that selected response only shows to what extent their students have learnt what the instructors have taught. Consequently, the instructors underestimate their demand to check whether their students can produce by using what they have learned in the English class.

Most instructors prefer selecting questions among the available ones though they acknowledge the importance of writing original questions in the study. Yet, preparing original questions is time-consuming and labor-intensive, therefore, the fear of possible increase in the instructors' workload cause the instructors to be dependent on their course and use the exercises in the course books without hesitation. In addition to the

course books, these instructors tend to use the questions on the Internet and their previous exams by assessing the questions and exams, which they consider more time-saver, more teacher-friendly and easier than writing original questions.

Most instructors in the study favor scoring selected response exams because they think selected response is easy to show the weights given to skills, grammar and listening in the course books. They also like grading selected response exams since all they need to is counting the number of their students' correct answers and multiply the number of the correct answers with the points given.

The concern about workload has caused the instructors to decrease the number of the components in the grading systems in the study. As some instructors seem to relate practicality to saving time in the study, they do not use assessment data to enhance their students' learning and their instruction though they know they should.

In addition to practicality, being critical shapes the practical dimension of language assessment literacy. The instructors in the study have developed their assessment knowledge mainly based on experience. Experience seems to help the instructors to acquire critical thinking to be used in assessment and evaluation in the study as stated in Lam (2015) and O'Loughlin (2013).

Critical thinking influences the instructors in developing appropriate assessments and making the exams valid and reliable in the study. Even if the instructors prefer choosing questions among the available ones, they develop their exams based on certain criteria that they have developed. They select words, listening audio and reading passages after the instructors assess and evaluate them depending on their criteria. When the instructors write questions, they use several critical thinking strategies (e.g., brainstorming, outlining and self-assessment).

The instructors have developed their own understanding of validity, reliability and measurement error though only three instructors have effective pre-service assessment training in the study. They apply several ways that they have learned as a result of experience to make the exams valid, reliable and free from measurement errors. The basis of the ways the instructors use for validity, reliability and avoiding measurement error is content validity because the instructors believe that if they make the exams content valid, the exams are valid, reliable and free from measurement error.

The instructors also determine whether their assessment results are valid or invalid mainly depending on content validity. Similarly, content validity, together with reliability, helps the instructors to check the consistency of their assessment results and interpretations. These findings indicate that the instructors are critical in different aspects of language assessment in this paragraph as they implement these aspects depending on their understanding in the study.

Besides practicality and critical thinking, being attentive influences the practical dimension of language assessment literacy in this study. The instructors do everything twice in grading the exams and announcing their grades in the study. They also administer their exams in a very careful way in order not to invalidate assessment results. In addition, the instructors have developed their understanding of ethical, legal and professional assessment practices mainly based on attentiveness because the ethical, legal and professional assessment practices that the instructors have mentioned require the instructors to be very careful to avoid making assessments invalid, unreliable and inconsistent in the study.

6.3. Implications of the Study

The findings of the study have indicated that the Turkish EFL instructors having graduated from ELL, EL and ACL departments perceived pedagogic formation assessment training as theoretical and numerical. The perception caused the instructors not to learn and use the knowledge of language assessment in assessment and evaluation. In addition, the findings have showed that though the Turkish EFL instructors having graduated from ELT department found pre-service language assessment training effective, they improved themselves through practicing the theory in real context. Finally, the findings have revealed that the instructors benefitted from different sources (e.g., peer assessment, peer feedback, observation, assessment experiences as a student, studying for exams and assessing and evaluating their students) to develop the knowledge of assessment.

All the findings have indicated have revealed a need to develop a language assessment literacy professional development program. The program should focus on the basic principles of language assessment suggested by Brown (2014): validity,

reliability, practicality, authenticity and washback. The basic principles should be explained in a clear and understandable way and supported with real examples. The professional development program should not be detailed and elaborated in statistics because the instructors in the study did not apply any statistical analysis to their assessment and evaluation practices. The possible participants of the professional development program should be given opportunity to assess and evaluate the basic principles in real language exams prepared by language teachers. The possible participants should also be shown the principles of preparing selected response, constructed response, performance assessment and personal communication and using how to score and grade different assessment methods as well as interpret and use assessment data. In addition, the program should enable the possible participants to apply the principles to their exams, assess their colleagues' exams, give feedback to their colleagues, evaluate the peer feedback and discuss the feedback with their colleagues. The program should cover administration, ethics and legality in assessment and evaluation. While doing so, the program should allow the possible participants to share what they have known with each other and compare the previous knowledge with the new knowledge they will learn in the program, so the participants can obtain positive experience as the findings of the present study have showed the instructors in the study developed their assessment knowledge based on experience. The professional development can be face-to-face and online.

Another implication of the study is that a national language assessment course book can be prepared for pre-service assessment training course and self-study. The course book should balance the practical aspect of language assessment and the theoretical aspect. The course book can be structured based on the suggestions made for the language assessment literacy professional development program.

6.4. Strengths of the Study

When the literature was reviewed, it was found out that language assessment literacy is a new term which has been studied for the last 10 years. Most of the studies related to language assessment literacy aimed at revealing the structures of pre-service language assessment training, the effects of professional development on language teachers and their knowledge base, assessment beliefs and practices and need for extra

training. None of those studied directly focused on how the language teachers implemented their language assessment literacy in their classes. However, the present study shows how language teachers implement it in their classes by providing first-hand data.

The studies in the related literature in Turkey are pre-service training, assessment knowledge base, assessment beliefs, assessment practices and exams. They do not indicate how language assessment literacy is implemented in language classes. Yet, this study helps to understand the implementation of language assessment literacy by Turkish EFL instructors in their English classes.

The participants of the study have different educational backgrounds. They graduated from ELT, ELL, EL and ACL departments of different Turkish universities and the ELT department of the faculty of the open university. As known, these departments, together with the department of ETI, provide the Turkish education system with English language teachers and instructors to teach English at the various levels of the education system. This study enables to understand how an English language teacher graduating from one of these departments has improved himself/herself in assessment and evaluation and how he/she implements his/her assessment and evaluation in his/her classes by dealing with each participant individually.

Finally, the studies in the literature indicate how the external factors like workload and the number of the students and the internal factors like assessment beliefs and previous experiences may affect language teachers, but they cannot show how language teachers are affected by the external and internal factors in terms of the seven sub-components of language assessment literacy. Despite this, this study presents how the external and internal factors influence language teachers when they assess and evaluate their students in a clear and detailed way.

6.5. Limitations of the Study

The first limitation is that this study was carried out in a specific context, so its results cannot be transferred to other contexts. Therefore, it cannot explain how other EFL instructors implement their assessment literacy in different institutions. The second limitation is time. This study lasted five months and presented its findings based on the

data collected during five months because it was very difficult for one person to observe and interview with eight participants, to transcribe the data collected, to analyze the data and to report the analysis. The third limitation is that the sample is small due to the nature of a qualitative study. The fourth limitation is that the study does not include any quantitative data to provide more concrete results, yet four observations and think-aloud protocol were used to compensate the lack of quantitative data in the study. Finally, the present study was based on the seven standards of assessment developed by AFT and its partner organizations (1990). This can be considered as a limitation because there are other professional organizations focusing and dealing with educational assessment and/or language assessment and the organizations can have different standards of assessment. The researcher did not depend only on the standards of AFT and its partner organizations in the study. He also searched and worked on the standards for, codes of and responsibilities for effective assessment prepared by NCME, JCTP, MAC, ACA, ITC and MONE. The seven standards of AFT and its partner organizations were the basis of such documents prepared by the other organizations. Therefore, the standards of AFT and its partner organizations were used in this study.

6.6. Suggestions for Further Research

The same research methodology and data collection instruments can be used in different institutions with different participants. The data from such studies can be compared and contrasted with each other including this study's data. Thus, a general and unique framework for language assessment literacy for Turkish EFL teachers/instructors can be formed. In addition, based on such data, pre-service, pedagogical formation and in-service training programs can be prepared to improve Turkish EFL teachers'/instructors' language assessment literacy and their results can be presented through articles and presentations in the international and national platforms. In addition, mixed-methods longitudinal studies can be made by several researchers on the implementation of language assessment literacy in the language class so that the limitations (e.g., lack of quantitative data, time and small sampling) can be overcome. Such studies can help the researchers to have generalizable results.

REFERENCES

- Abell, S. K. & Siegel, M. A. (2011). Assessment literacy: What science teachers need to know and be able to do. In D. Corrigan, J. Dillon, & R. Gunstone (Eds.), *The professional knowledge base of science teaching* (pp. 205-221). Springer Business+Media Media B.V.
- Adanalı, K. & Dođanay, A. (2010). Beşinci sınıf sosyal bilgiler öğretimini alternatif ölçme-değerlendirme etkinlikleri açısından değerlendirilmesi. *Çukurova Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 19(1), 271-292.
- Akdağ, G. & Ekmekçi, S. (2015). Fen ve teknoloji öğretmenlerinin ölçme-değerlendirmeye ilişkin yeterlik algıları ve görüşleri. *Route Educational and Social Science Journal*, 2(3), 253-273.
- Aksu Ataç, B. (2012). Foreign language teachers' attitude toward authentic assessment in language teaching. *The Journal of Language and Linguistic Studies*, 8(2), 7-19.
- Alas, E. & Liiv, S. (2014). Assessment literacy of national examination interviewers and rater-experience with the CEFR. *Eesti Rakenduslingvistika Ühingu Aastaraamat*, 10, 7-22.
- Alduais, A. M. S. (2012). An account of approaches to language testing. *International Journal of Academic Research in Progressive Education and Development*, 1(4), 203-208.
- Alkharusi, H. (2011a). Self-perceived assessment skills of pre-service and in-service teachers. *Jurnal Pendidikan Malaysia*, 36(2), 9-17.
- Alkharusi, H. (2011b). Psychometric properties of the teacher assessment literacy questionnaire for preservice teachers in Oman. *Procedia – Social and Behavioral Sciences*, 29, 1614-1624.
- Alkharusi, H. (2011c). An analysis of the internal and external structure of the teacher assessment literacy questionnaire. *International Journal of Learning*, 18, 515-528.

- Alkharusi, H. (2011d). Teachers' classroom assessment skills: Influence of gender, subject area, grade level, teaching experience and in-service assessment training. *Journal of Turkish Science Education*, 8(2), 39-48.
- Alkharusi, H., Aldhafri, S., Alnabhani, H., & Alkalbani, M. (2012). Educational assessment attitudes, competence, knowledge and practices: An exploratory study on Muscat teachers in the Sultanate of Oman. *Journal of Education and Learning*, 1(2), 217-217-232.
- Altun, A. & Gelbal, S. (2014). Öğretmenlerin kullandıkları ölçme ve değerlendirme yöntem veya araçlarının ikili karşılaştırma yöntemiyle belirlenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 5(1), 1-11.
- American Counseling Association. (2003). *Standards for qualifications of test users*. Retrieved from <http://aac.ncat.edu/documents/Standards%20for%20Qualifications%20of%20Test%20Users.DOC>
- American Federation of Teachers, National Council on Measurement in Education, & National Education Association. (1990). *Standards for teacher competence in educational assessment of students*. Retrieved from <http://buros.org/standards-teacher-competence-educational-assessment-students>
- Ataman, M. & Kabapınar, Y. (2012). Sosyal bilgiler (4-5. sınıf) programlarındaki ölçme değerlendirme yöntemlerinin kullanılma nedenleri ve uygulamalarının yeterliliği. *Amasya Üniversitesi Eğitim Fakültesi Dergisi*, 1(1), 94-114.
- Atıkol, R. (2008). *In-service English teachers' opinions of assessment and evaluation of young learners: Portfolio assessment as an alternative* (Unpublished master thesis). Çanakkale Onsekiz Mart University, Çanakkale, Turkey.
- Aydoğmuş, A. & Çoşkun Keskin, S. (2012). Sosyal bilgiler öğretmenlerinin süreç odaklı ölçme değerlendirme araçlarını kullanma durumları: İstanbul ili örneği. *Mersin Üniversitesi Eğitim Fakültesi Dergisi*, 8(2), 110-123.
- Bailey, K. M. & Brown, J. D. (1996). Language testing courses: What are they? In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 236-256). Clevedon, UK: Multilingual Matters.

- Baker, B. A., Tsushima, R., & Wang, S. (2014). Investigating language assessment literacy: Collaboration between assessment specialists and Canadian admissions officers. *CercleS*, 4(1), 137-157. doi: 10.1515/cercles-2014-0009
- Black, P., Harrison, C., Lee, C., Marshall B., & William, D. (2003). *Assessment for learning: Putting it into practice*. Maidenhead: Open University Press.
- Boraie, D. (2012). *Formative assessment vs. summative assessment: Does it matter?* Retrieved from <http://newsmanager.commpartners.com/tesolc/issues/2012-09-01/3.html>
- Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal*, 9(2), 27-40. doi: 10.3316/QRJ0902027
- Boyd, E. (2015). *Assessment literacy for teachers: How to identify and write a good test*. In G. Pickering & P. Gunashekar (Eds.), *Innovation in English language teacher education*. Paper presented at the 4th International Teacher Education Conference, Hyderabad, India (pp. 134-140). British Council.
- Bracey, G. W. (2000). *Thinking about tests and testing: A short premier in "assessment literacy"*. Retrieved from <http://www.aypf.org/publications/braceyrep.pdf>
- Braney, B. (2011, April). *An examination of fourth-grade teachers' assessment literacy and its relationship to students' reading achievement*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Brindley, G. (2001). Language assessment and professional development. In C. Elder, A. Brown, E. Grove, K. Hall, N. Iwashita, T. Lumley, ... K. O'Loughlin (Eds.), *Experimenting with uncertainty. Essays in honour of Alan Davies* (pp. 126-136). Cambridge, UK: Cambridge University Press.
- Brindley, G. (2008). Educational reform and language testing. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed.) (pp. 365-378). Springer Science+Business Media LLC.
- Broadfoot, P. M. (2005). Dark alleys and blind bends: Testing the language of learning. *Language Testing*, 22(2), 123-141. doi: 10.1191/0265532205lt302oa

- Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. Turkey: Pearson Education Inc.
- Brown, J. D. (2005). Characteristics of sound qualitative research. *Shiken: JALT Testing & Evaluation SIG Newsletter*, 9(2), 31-33.
- Brown, J. D. & Bailey, K. M. (2008). Language testing courses: What are they in 2007? *Language Testing*, 25(3), 349-383. doi: 10.1177/0265532208090157
- Bugel, M. J. (2011). “*Out of the Mouths of Sibs*”... *A phenomenological study of the experience of being a well school-age child with a traumatic injury* (Unpublished doctoral dissertation). Seton Hall University, New Jersey, America.
- Büyükkarıcı, K. (2014). Assessment beliefs and practices of language teachers in primary education. *International Journal of Instruction*, 7(1), 107-120.
- Büyükkarıcı, K. (2016). Identifying the areas for English language teacher development: A study of assessment literacy. *Pegem Eğitim ve Öğretim Dergisi*, 6(3), 333-346.
- Campbell, C., Murphy, J. A., & Holt, J. K. (2002, October). *Psychometric analysis of an assessment literacy instrument: Applicability to preservice teachers*. Paper presented at the annual meeting of the Mid-Western Educational Research Association, Columbus, OH.
- Cansız Aktaş, M. & Baki, A. (2013). Yeni ortaöğretim matematik öğretim programının ölçme değerlendirme boyutuyla ilgili öğretmen görüşleri. *Kastamonu Eğitim Dergisi*, 21(1), 203-222.
- Carcary, M. (2009). The research audit trail – enhancing trustworthiness in qualitative inquiry. *The Electronic Journal of Business Research Methods*, 7(1), 11-24.
- Chan, Y-C. (2008). Elementary school EFL teachers’ beliefs and practices of multiple assessments. *Reflections on English Language Teaching*, 7(1), 37-62.
- Cheng, L. (2008). Washback, impact and consequences. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed.) (pp. 349-364). Springer Science+Business Media LLC.

- Cohen, L., Manion, L., & Morrison, K. (2005). *Research methods in education* (5th ed.). New York, NY: RoutledgeFalmer.
- Colker, A. M. (n.d.). *Developing interviews: Preparing an interview protocol*. Retrieved from http://oerl.sri.com/module/mod6/m6_p1.html
- Creswell, J. W. (2007). *Qualitative inquiry & research design: Choosing among five approaches* (2nd ed.). Thousand Oaks, California, the United States of America: Sage Publications.
- Davidheiser, S. A. (2013). *Identifying areas for high school teacher development: A study of assessment literacy in the Central Bucks School District* (Doctoral dissertation). Retrieved from <https://idea.library.drexel.edu/islandora/object/idea:4170>. (Accession number: 4170).
- Davidson, C. (2009). Transcription: Imperatives for qualitative research. *International Journal of Qualitative Methods*, 8(2), 35-52.
- Davies, A. (2008). Textbook trends in teaching language testing. *Language Testing*, 25(3), 327-347. doi: 10.1177/0265532208090156
- Davison, C. (2004). The contradictory culture of teacher-based assessment: ESL teacher assessment practices in Australian and Hong Kong secondary schools. *Language Testing*, 21(3), 305-334. doi: 10.1191/0265532204lt286oa
- Davison, C. & Leung, C. (2009). Current issues in English language teacher-based assessment. *TESOL Quarterly*, 43(3), 393-415.
- DeLuca, C., Chavez, T., Bellara, A., & Cao, C. (2013). Pedagogies for preservice assessment education: Supporting teacher candidates' assessment literacy development. *The Teacher Educator*, 48(2), 128-142. doi: 10.1080/08878730.2012.760024
- DeLuca, C. & Klinger, D. A. (2010). Assessment literacy development: Identifying gaps in teacher candidates' learning. *Assessment in Education: Principles, Policy & Practice*, 17(4), 419-438. doi:10.1080/0969594X.2010.516643

- DeLuca, C., Luu, K., Sun, Y., & Klinger, D. A. (2012). Assessment for learning in the classroom: Barriers to implementation and possibilities for teacher professional development. *Assessment Matters*, 4, 5-29.
- Dörnyei, Z. (2011). *Research methods in applied linguistics*. Oxford, UK: Oxford University Press.
- Duong, T. M., Pham, T. T. H., Thai, H. L. T. (n.d.). *Building an assessment competence framework for pre-service and in-service teachers in Vietnam*. Retrieved from http://www.ibrarian.net/navon/paper/Building_an_Assessment_Compotence_Framework_for_P.pdf?paperid=17972402
- Eğri, G. (2006). *Coğrafya öğretmenlerinin ölçme değerlendirme yapabilme yeterliliği* (Unpublished master thesis). Gazi Üniversitesi, Ankara, Türkiye.
- Engelsen, K. S. & Smith, K. (2014). Assessment literacy. In C. Wyatt-Smith, V. Klenoswki, & P. Colbert (Eds.), *Designing assessment for quality learning* (pp. 91-107). Springer Science+Business Media Dordrecht.
- English as a Second Language Council of the Alberta Teachers' Association. (2010). *Understanding ESL learners: Assessment*. Retrieved from <http://www.teachers.ab.ca/SiteCollectionDocuments/ATA/Publications/Specialist-Councils/ESL-3-4%20Assessment.pdf>
- Fairchild, A. J. (n.d.). *Instrument reliability and validity: Introductory concepts and measures*. Retrieved from http://www.jmu.edu/assessment/wm_library/Reliability_validity.pdf
- Fan, Y-C., Wang, T-H., & Wang, K-H. (2011). A Web-based model for developing assessment literacy of secondary in-service teachers. *Computers & Education*, 57, 1727-1740.
- Fonteyn, M. E., Kuipers, B., & Grobe, S. J. (1993). A description of think aloud method and protocol analysis. *Qualitative Health Research*, 3(4), 430-441.
- Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly*, 9(2), 113-132. doi: 10.1080/15434303.2011.642041

- Gökçe, Ö. F. (2014). *A comparison of EFL teachers' perception of formative assessment in public and private schools* (Unpublished master thesis). Çağ University, Mersin, Turkey.
- Gönen, K. & Akbarov, A. (2015). Instructors' principles and practices of classroom-based language assessment in higher education in Turkey. *Journal of European Education*, 5(3), 28-38.
- Gotch, C. M. & French, B. F. (2014). A systematic review of assessment literacy measures. *Educational Measurement: Issues and Practice*, 33(2), 14-18.
- Gottheiner, D. M. & Siegel, M. A. (2012). Experienced middle school science teachers' assessment literacy: Investigating knowledge of students' conceptions in genetics and ways to shape instruction. *Journal of Science Teacher Education*, 23, 531-557. doi: 10.1007/s10972-012-9278-z
- Green, S. (2002). *Criterion referenced assessment as a guide to learning the importance of progression and reliability*. Retrieved from <http://www.cambridgeassessment.org.uk/images/109693-criterion-referenced-assessment-as-a-guide-to-learning-the-importance-of-progression-and-reliability.pdf>
- Guerin, E. M. C. (2010). *Initial findings from a pilot Italian study of foreign language teachers' stated language assessment knowledge-base and needs*. Paper presented at the Lancaster University Postgraduate Conference in Linguistics & Language Teaching, Lancaster, the United Kingdom.
- Gül, E. (2011). *İlköğretim öğretmen adaylarının ölçme-değerlendirme okuryazarlığı ve ölçme-değerlendirmeye ilişkin tutumlarının belirlenmesi* (Unpublished master thesis). Fırat Üniversitesi, Elazığ, Türkiye.
- Hakim, B. (2015). English language teachers' ideology of ELT assessment literacy. *International Journal of Education & Literacy Studies*, 3(4), 42-48.
- Hamp-Lyons, L. (2017). Language assessment literacy for language-oriented assessment. *Papers in Language Testing and Assessment*, 6(1), 88-111.

- Han, T. & Kaya, H. İ. (2014). Turkish EFL teachers' assessment preferences and practices in the context of constructivist instruction. *Journal of Studies in Education*, 4(1), 77-93.
- Hancock, D. R. & Algozzine, B. (2006). *Doing case study research: A practical guide for beginning researchers*. New York, NY: Teachers College Press.
- Hasselgreen, A., Carlsen, C., & Helness, H. (2004). *European survey of language testing and assessment needs. Report: Part one – general findings*. Retrieved from <http://www.ealta.eu.org/documents/resources/survey-report-pt1.pdf>
- Hatipoğlu, Ç. (2010, Winter). Summative evaluation of an English language testing and evaluation course for future English language teachers in Turkey. *ELTED*, 13, 40-51.
- Hatipoğlu, Ç. (2015a). English language testing and evaluation (ELTE) training in Turkey: Expectations and needs of pre-service English language teachers. *ELT Research Journal*, 4(2), 111-128.
- Hatipoğlu, Ç. (2015b, May). *Diversity in language testing and assessment literacy of language teachers in Turkey*. Paper presented at the 3rd ULEAD Congress, International Congress on Applied Linguistics: Current Issues in Applied Linguistics, Çanakkale, Turkey.
- Hatipoğlu, Ç. & Erçetin, G. (2016). Türkiye'de yabancı dilde ölçme ve değerlendirme eğitiminin geçmişi ve bugünü. In S. Akcan & Y. Bayyurt (Eds), *3. ulusal yabancı dil eğitimi kurultayı: Türkiye'deki yabancı dil eğitimi üzerine görüş ve düşünceler 23-24 Ekim 2014, konferanstan seçkiler* (pp. 72-89). İstanbul, Turkey: Boğaziçi University.
- Herrera, L. & Macias, D. (2015). A call for language assessment literacy in the education and development of teachers of English as a foreign language. *Colombian Applied Linguistics Journal*, 17(2), 302-312.
- Hidri, S. (2015). Conceptions of assessment: Investigating what assessment means to secondary and university teachers. *Arab Journal of Applied Linguistics*, 1(1), 19-43.

- Hill K. (2017). Understanding classroom-based assessment practices: A precondition for teacher assessment literacy. *Papers in Language Testing and Assessment*, 6(1), 1-17.
- Huai, N., Braden, J. P., White, J. L., & Elliott, S. (2006). Effect of an Internet-based professional development program on teachers' assessment literacy for all students. *Teacher Education and Special Education*, 29(4), 244-260.
- Huhta, A., Hirvela, T., & Banerjee, J. (2005). *European survey of language testing and assessment needs. Report: Part two – regional findings*. Retrieved from http://users.jyu.fi/~huhta/ENLTA2/First_page.htm
- Inbar-Lourie, O. (2008a). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing*, 25(3), 385-402. doi: 10.1177/0265532208090158
- Inbar-Lourie, O. (2008b). Language assessment culture. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed.) (pp. 285-299). Springer Science+Business Media LLC.
- Inbar-Lourie, O. (2013). Language assessment literacy. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 1-8). Blackwell Publishing Ltd.
- International Test Commission. (2001). International guidelines for test use. *International Journal of Testing*, 1(2), 93-114.
- International Test Commission. (2005). *International guidelines for translating and adapting test*. Retrieved from http://www.intestcom.org/files/guideline_test_adaptation.pdf
- International Test Commission. (2012). *International guidelines on quality control in scoring, test analysis and reporting of test scores*. Retrieved from https://www.nite.org.il/files/QC_Guidelines.pdf
- International Test Commission. (2014). *The ITC guidelines on the security of tests, examinations and other assessments*. Retrieved from http://www.intestcom.org/files/guideline_test_security.pdf

- Izci, K. & Siegel, M. (2014, April). *Investigating high school chemistry teachers' assessment literacy in theory and practice*. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.
- Jacob, S. A., & Furgerson, S. P. (2012). Writing interview protocols and conducting interviews: Tips for students new to the field of qualitative research. *The Qualitative Report, 17*(6), 1-10.
- Jannati, S. (2015). ELT teachers' language assessment literacy: Perceptions and practices. *The International Journal of Research in Teacher Education, 6*(2), 26-37.
- Jeong, H. (2013). Defining assessment literacy: Is it different for language testers and non-language testers? *Language Testing, 30*(3), 345-362.
- Jin, Y. (2010). The place of language testing and assessment in the preparation of foreign language teachers in China. *Language Testing, 27*(4), 555-584. doi: 10.1177/0265532209351431
- Joint Committee on Testing Practices. (2002). *Code of fair testing practices in education*. Retrieved from <http://apa.org/science/programs/testing/fair-code.aspx>
- Kahl, S. R., Hofman, P., & Bryant, S. (2013). *Assessment literacy standards and performance measures for teacher candidates and practicing teachers*. Retrieved from https://secure.aacte.org/apps/rl/res_get.php?fid=1364&ref=rl
- Karaman, P. & Şahin, Ç. (2014). Öğretmen adaylarının ölçme değerlendirme okuryazarlıklarının belirlenmesi. *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi, 15*(2), 175-189.
- Khadijeh, B. & Amir, R. (2015). Importance of teachers' assessment literacy. *International Journal of English Language Education, 3*(1), 139-146.
- Kiomrs, R., Abdolmehdi, R., & Naser, R. (2011). On the interaction of test washback and teacher assessment literacy: The case of Iranian EFL secondary school teachers. *English Language Testing, 4*(1), 156-161.
- Kitzinger, J. (1995). Introducing focus groups. *BMJ, 311*, 299-302.

- Kırkgöz, Y. & Ağçam, R. (2012). Investigating the written assessment practices of Turkish teachers of English at primary education. *The Journal of Language and Linguistic Studies*, 8(2), 119-136.
- Koh, K. H. (2011). Improving teachers' assessment literacy through professional development. *Teaching Education*, 22(3), 255-276. doi:10.1080/10476210.2011.593164
- Koh, K. H. & Velayutham, R. L. (n.d.). *Improving teachers' assessment literacy in Singapore schools: An analysis of teachers' assessment tasks and student work*. Retrieved from https://www.nie.edu.sg/files/NIE_research_brief_09_002.pdf
- Köksal, D. (2004). Assessing teachers' testing skills in ELT and enhancing their professional development through distance learning on the net. *Turkish Online Journal of Distance Education-TOJDE*, 5(1).
- Krueger, R. A. (2002). *Designing and conducting focus group interviews*. Retrieved from <http://www.eiu.edu/ihec/Krueger-FocusGroupInterviews.pdf>
- Kuran, K. & Kanatlı, F. (2009). Alternatif ölçme ve değerlendirme teknikleri konusunda sınıf öğretmenlerinin görüşlerinin değerlendirilmesi. *Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 6(12), 209-234.
- Lam, R. (2015). Language assessment training in Hong Kong: Implications for language assessment literacy. *Language Testing*, 32(2), 169-197. doi: 10.1177/0265532214554321
- Leaph, K., Channy, M., & Chan, N. K. (2015). Cambodian ELT university practitioners' use of standardized tests for practice and assessment. *Language Education in Asia*, 6(1), 4-16.
- Leighton, J. P., Gokiart, R. J., Cor, M. K., & Heffernan, C. (2010). Teacher beliefs about the cognitive diagnostic information of classroom- versus large-scale tests: Implications for assessment literacy. *Assessment in Education: Principles, Policy & Practice*, 17(1), 7-21. doi: 10.1080/09695940903565362
- Leong, W. S. & Tan, K. (2014). What (more) can and should, assessment do for learning? Observations from 'successful learning context' in Singapore. *The Curriculum Journal*, 25(4), 593-619. doi:10.1080/09585176.2014.970207

- Leung, C. & Lewkowicz, J. (2006). Expanding horizons and unresolved conundrums: Language testing and assessment. *TESOL Quarterly*, 40(1), 221-234.
- Levy-Vered, A. & Nasser-Abu Alhija, F. M. (2014, April). *Modelling assessment literacy of beginning teachers: The contribution of training and conceptions*. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.
- Lian, L. H., Yew, W. T., & Meng, C. C. (2014). Enhancing Malaysian teachers' assessment literacy. *International Education Studies*, 7(10), 74-81.
- Lincoln, Y. S. & Guba, E. G. (1985). *Naturalistic inquiry*. Newbury Park, CA: Sage Publications.
- Lomax, R. G. (1996). On becoming assessment literate: An initial look at preservice teachers' beliefs and practices. *The Teacher Educator*, 31(4), 292-303. doi: 10.1080/08878739609555122
- Lukin, L. E., Bandalos, D. L., Eckhout, T. J., & Mickelson, K. (2004, Summer). Facilitating the development of assessment literacy. *Educational Measurement: Issues and Practice*, 26-32.
- Lynch, B. K. (2001). Rethinking assessment from a critical perspective. *Language Testing*, 18(4), 351-372. doi: 10.1177/026553220101800403
- Lysaght, Z. (2015). Assessment for learning and for self-regulation. *The International Journal of Emotional Education*, 7(1), 20-34.
- Mahapatra, S. K. (2013). Assessment literacy. A panacea for many problems in language assessment. *FORTELL*, 27, 9-11.
- Mahapatra, S. K. (2016). Tracking the development of teachers' language assessment literacy: A case study. In G. Pickering & P. Gunashekar (Eds.), *Ensuring quality in English language teacher education* (pp. 106-114). New Delhi, India: British Council.
- Malone, M. E. (2008). Training in language assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed.) (pp. 225-239). Springer Science+Business Media LLC.

- Malone, M. E. (2011). *Assessment literacy for language educators*. Retrieved from <http://www.cal.org/index.php/content/download/1516/15923/file/AssessmentLiteracyforLanguageEducators.pdf>
- Malone, M. E. (2013). The essentials of assessment literacy: Contrasts between testers and users. *Language Testing*, 30(3), 329-344. doi: 10.1177/0265532213480129
- McGee, J. & Colby, S. (2014). Impact of an assessment course on teacher candidates' assessment literacy. *Action in Teacher Education*, 36(5-6), 522-532. doi: 10.1080/01626620.2014.977753
- McDowell, L., Sambell, K., Bazin, V., Penlington, R., Wakelin, D., Wickes, H., & Smailes, J. (2006). *Assessment for learning: Current practice exemplars for the Centre for Excellence in Teaching and Learning*. Newcastle, UK: Northumbria University.
- McNamara, T. (2001). Language assessment as social practice: Challenges for research. *Language Testing*, 18(4), 333-349. doi: 10.1177/026553220101800402
- Mede, E. & Atay, D. (2017). English language teachers' assessment literacy: The Turkish context. *Ankara Üniversitesi TÖMER Dil Dergisi*, 168(1), 43-60.
- Mercurio, A. (2013). *Assessment and learning: Building teachers' assessment literacy*. Retrieved from http://www.iaea.info/documents/paper_5bc1b798.pdf
- Mertler, C. A. (2000). Teacher-centered fallacies of classroom assessment validity and reliability. *Mid-Western Educational Researcher*, 13(4), 29-35.
- Mertler, C. A. (2003, October). *Preservice versus inservice teachers' assessment literacy: Does classroom experience make a difference?* Paper presented at the annual meeting of the Mid-Western Educational Research Association, Columbus, OH.
- Mertler, C. A. (2009). Teachers' assessment knowledge and their perceptions of the impact of classroom assessment professional development. *Improving Schools*, 12(2), 101-113. doi:1177/1365480209105575
- Mertler, C. A. & Campbell, C. (2005, April). *Measuring teachers' knowledge & application of classroom assessment concepts: Development of the Assessment*

Literacy Inventory. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Michigan Assessment Consortium. (2013). *Assessment literacy standards*. Retrieved from

<http://www.michiganassessmentconsortium.org/sites/default/files/Assessment%20Literacy%20Standards-Students.docx>

Milli Eğitim Bakanlığı. (n.d.). *Öğretmenlik mesleği genel yeterlikleri*. Retrieved from http://otmg.meb.gov.tr/Yeterlik_surec.html

Montee, M., Bach, A., Donovan, A., & Thompson, L. (2013, Spring). LCTL teachers' assessment knowledge and practices: An exploratory study. *Journal of the National Council of Less Commonly Taught Languages*, 14, 1-32.

Munoz, A. P., Palacio, M., & Escobar, L. (2012). Teachers' beliefs about assessment in an EFL context in Colombia. *PROFILE*, 14(1), 143-158.

National Council on Measurement in Education. (1995). *Code of professional responsibilities in educational measurement*. Retrieved from http://www.niu.edu/assessment/manual/_docs/EthicsCode.pdf

Newfields, T. (2006). *Teacher development and assessment literacy*. Paper presented at the Proceedings of the 5th Annual JALT Pan-SIG Conference, Shizuoka, Japan (pp. 48-73). Tokai University College of Marine Science.

Nier, V. C., Donovan, A. E., & Malone, M. E. (2013). *Promoting assessment literacy for language instructors through an online course*. Retrieved from http://www.cal.org/ecolt/pdfs/AB_Poster_10-21-2013_FINAL.pdf

North Central Regional Educational Laboratory. (n.d.). *Indicator: Assessment*. Retrieved from <http://www.ncrel.org/engage/framework/pro/literacy/prolitin.htm>

O'Loughlin, K. (2009). *Developing the assessment literacy of IELTS test users in higher education* (Report No. 13). Retrieved from http://www.ielts.org/pdf/vol13_report5.pdf

O'Loughlin, K. (2013). Developing the assessment literacy of university proficiency test users. *Language Testing*, 30(3), 363-380. doi: 10.1177/0265532213480336

- O'Sullivan, R. G. & Johnson, R. L. (1993, April). *Using performance assessments to measure teachers' competence in classroom assessment*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.
- Örsdemir, E. (2010). *Alternative assessment in Turkish primary EFL classrooms: An investigation into the performance tasks* (Unpublished master thesis). Çukurova University, Adana, Turkey.
- Öz, H. (2014). Turkish teachers' practices for *assessment for learning* in the English as a foreign language classroom. *Journal of Language Teaching and Research*, 5(4), 775-785.
- Öz, S. & Atay, D. (2017). Turkish EFL instructors' in-class language assessment literacy: Perceptions and practices. *ELT Research Journal*, 6(1), 25-44.
- Özer, B. & Karaoğlu, A. (2017). Fen ve teknoloji derslerinde kullanılan tamamlayıcı ölçme-değerlendirme yöntemlerinin incelenmesi. *Uluslararası Türk Eğitim Bilimleri Dergisi*, 5(3), 129-141.
- Paterno, J. (2001). *Measuring success: A glossary of assessment terms*. Retrieved from <http://www.angelfire.com/wa2/buildingcathedrals/measuringsuccess.html>
- Patton, M. Q. (2002). *Qualitative evaluation and research methods* (3rd Ed.). Newbury Park: Sage Publications.
- Perry, M. L. (2013). *Teacher and principal assessment literacy* (Doctoral dissertation). Retrieved from <http://scholarworks.umt.edu/cgi/viewcontent.cgi?article=2410&context=etd>. (Accession number: 1391).
- Phelan, C. & Wren, J. (2005). *Exploring reliability in academic assessment*. Retrieved from <https://www.uni.edu/chfasoa/reliabilityandvalidity.htm>
- Pill, J. & Harding, L. (2013). Defining the language assessment literacy gap: Evidence from a parliamentary inquiry. *Language Testing*, 30(3), 381-402. doi: 10.1177/0265532213480337
- Pitney, W. A. (2004). Strategies for establishing trustworthiness in qualitative research. *Athletic Therapy Today*, 9(1), 26-28.

- Plake, B. S. (1993). Teacher assessment literacy: Teachers' competence in the educational assessment of students. *Mid-Western Educational Researcher*, 6(1), 21-27.
- Plake, B. S., Impara, J. C., & Fager, J. J. (1993). Assessment competencies of teachers: A national survey. *Educational Measurement: Issues and Practice*, 12(4), 10-39.
- Popham, W. J. (2004, September). All about accountability / why assessment illiteracy is professional suicide. *Educational Leadership*, 62(1), 82-83.
- Popham, W. J. (2006). All about accountability / needed: A dose of assessment literacy. *Educational Leadership*, 63(6), 84-85.
- Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory into Practice*, 48(1), 4-11. doi: 10.1080/00405840802577536
- Popham, W. J. (2011). Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator*, 46(4), 265-273. doi: 10.1080/08878730.2011.605048
- Popham, W. J. (2012). *Mastering assessment: A self-directed system for educators. Assessment bias: How to banish it* (2nd ed.). Boston, MA: Pearson Education, Inc.
- Quilter, S. M. & Gallini, J. K. (2000). Teachers' assessment literacy and attitudes. *The Teacher Educator*, 36(2), 115-131. doi:10.1080/08878730009555257
- Quitter, S. M. (1999). Assessment literacy for teachers: Making a case for the study of test validity. *The Teacher Education*, 34(4), 235-243. doi: 10.1080/08878739909555204
- Razavipour, K. (2014). Assessing assessment literacy: Insights from a high-stakes test. *RALS*, 4(1), 111-131.
- Rea-Dickins, P. (2004). Understanding teachers as agents of assessment. *Language Testing*, 21(3), 249-258.
- Rea-Dickins, P. (2006). Currents and eddies in the discourse of assessment: A learning-focused interpretation. *International Journal of Applied Linguistics*, 16(2), 164-188.

- Rea-Dickins, P. (2008). Classroom-based language assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed.) (pp. 257-271). Springer Science+Business Media LLC.
- Riazi, A. M. & Razavipour, K. (2011). (In) Agency of EFL teachers under the negative backwash effect of centralized tests. *International Journal of Language Studies*, 5(2), 123-142.
- Richardson McGee, J. & Colby, S. (2015, April). *Changes in teacher candidates' assessment literacy*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Riestenberg, K., Silvio, F. D., Donovan, A., & Malone, M. E. (2010, Fall). Development of a computer-based workshop to foster language assessment literacy. *Journal of the National Council of Less Commonly Taught Languages*, 9, 21-42.
- Rogers, W. T., Cheng, L., & Hu, H. (2007). ESL/EFL instructors' beliefs about assessment and evaluation. *Canadian and International Education/Education canadienne at interntionale*, 36(1), 39-61.
- Rogier, D. (2014). Assessment literacy: Building a base for better teaching and learning. *English Teaching Forum*, 3, 2-13.
- Rubin, H. J. & Rubin, I. S. (2005). *Qualitative interviewing: The art of hearing data* (2nd ed.). America: Sage Publications, Inc.
- Saad, M. R. B. M., Sardareh, S. A., & Ambarwati, E. K. (2013). Iranian secondary school EFL teachers' assessment beliefs and roles. *Life Science Journal*, 10(3), 1638-1647.
- Şahin, S. (2015, May). *Language testing and assessment (LTA) literacy of high school English language teachers in Turkey*. Paper presented at the 3rd ULEAD Congress, International Congress on Applied Linguistics: Current Issues in Applied Linguistics, Çanakkale, Turkey.
- Sarıçoban, A. (2011). A study on the English language teachers' preparation of tests. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 41, 398-410.

- Scarino, A. (2013). Language assessment literacy as self-awareness: *Understanding the role of interpretation in assessment and in teacher learning*. *Language Testing*, 30(3), 309-327. doi: 10.1177/0265532213480128
- Scarino, A. (2017). Developing assessment literacy of teachers of languages: A conceptual and interpretive challenge. *Papers in Language Testing and Assessment*, 6(1), 18-40.
- Seidman, I. (2006). *Interviewing as qualitative research* (3rd ed.). New York, NY: Teachers College Press.
- Sellan, R. (2017). Developing language assessment literacy in Singapore: How teachers broaden English language learning by expanding assessment constructs. *Papers in Language Testing and Assessment*, 6(1), 64-87.
- Sever, I. & İflazoğlu Saban, A. (2015). Öğretim elemanlarının ölçme ve değerlendirme yeterlik algılarının belirlenmesi. *Uludağ Üniversitesi Eğitim Fakültesi Dergisi*, 28(2), 173-204.
- Sezer, C. (2010). *Okul öncesi öğretmenlerinin ölçme değerlendirmeyi kullanma düzeylerinin belirlenmesi* (Unpublished master thesis). Çanakkale Onsekiz Mart Üniversitesi, Çanakkale, Türkiye.
- Shenton, A. K. (2004). Strategies for ensuring trustworthiness in qualitative research. *Education for Information*, 22, 63-75.
- Shohamy, E., Inbar-Lourie, O., & Poehner, M. (2008). *Investigating assessment perceptions and practices in the advanced foreign language classroom* (Report No. 1108). University Park, PA: Center for Advanced Language Proficiency Education and Research.
- Siegel, M. A. & Wissehr, C. (2011). Preparing for the plunge: Pre-service teachers' assessment literacy. *Journal of Science Teacher Education*, 22, 371-391. doi:10.1007/s10972-011-9231-6
- Smith, C. D., Worsfold, K., Davies, L., Fisher, R., & McPhail, R. (2013). Assessment literacy and student learning: The case for explicitly developing students' assessment literacy. *Assessment & Evaluation in Higher Education*, 38(1), 44-60. doi: 10.1080/02602938.2011.598636

- Stakes, R. E. (1995). *The art of case study research*. The United States of America: Sage Publications, Inc.
- Stiggins, R. (2005a, September). *Assessment for learning defined*. Retrieved from <http://ati.pearson.com/downloads/afldefined.pdf>
- Stiggins, R. (2005b, December). From formative assessment to assessment FOR learning: A path to success in standards-based schools. *Phi Delta Kappan*, 87(04), 324-328.
- Stiggins, R. (2006, November/December). Assessment for learning: A key to motivation and achievement. *EDge*, 2(2), 3-19.
- Stiggins, R. (2007, May). Assessment through the student's eyes. *Educational Leadership*, 64(8), 22-26.
- Stiggins, R. (2014, October). Improve assessment literacy outside of schools too. *Kappan Magazine*, 96(2), 67-72.
- Stiggins, R. & Chappuis, J. (2006, Winter). What a difference a word makes: Assessment FOR learning rather than assessment OF learning helps students succeed. *Journal of Staff Development*, 27(1), 10-14.
- Stiggins, R. & Chappuis, S. (2005, October). Putting testing in perspective: It is for learning. *Measuring Up*, 16-20.
- Stiggins, R. & Popham, W. J. (n.d.). *Assessing students' affect related to assessment for learning: An introduction for teachers*. Retrieved from http://www.ccsso.org/Documents/2007/Assessing_Students_Affect_2007.pdf
- Stiggins, R. J. (1991). Assessment literacy. *The Phi Delta Kappan*, 72(7), 534-539.
- Stiggins, R. J. (1995). Assessment literacy for the 21st century. *The Phi Delta Kappan*, 77(3), 238-245.
- Stiggins, R. J. (1999). Assessment, Student Confidence and School Success. *The Phi Delta Kappan*, 81(3), 191-198.
- Stiggins, R. J. (2002, June). Assessment crisis: The absence of assessment FOR learning. *Phi Delta Kappan*, 1-10.

- Stiggins, R., Arter, J., Chappuis, J., & Chappuis, S. (2004). *Classroom assessment for student learning: Doing it right-using it well*. Assessment Training Institute.
- Tahmasbi, S. (2014). Scaffolding and artifacts. *Middle-East Journal of Scientific Research*, 19(10), 1378-1387.
- Tulgar, A. T. (2017). Selfie@ssessment as an alternative form of self-assessment at undergraduate level in higher education. *Journal of Language and Linguistics Studies*, 13(1), 321-335.
- Talib, R., Kamsah, M. Z., Ghafar, M. N. A., Zakaria, M. A. Z. M., & Naim, H. A. (2013). *T-assess: Assessment literacy for Malaysian teachers*. Paper presented at the International Conference on Assessment for Higher Education Across Domains and Skills, Kuala Lumpur.
- Taylor, L. (2009). Developing assessment literacy. In *Annual Review of Applied Linguistics* (pp. 21-36). The U.S.A.: Cambridge University Press.
- Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing*, 30(3), 403-412. doi: 10.1177/0265532213480338
- Tsagari, D. & Vogt, K. (2017). Assessment literacy of foreign language teachers around Europe: Research, challenges and future prospects. *Papers in Language Testing and Assessment*, 6(1), 41-63.
- Turner, III, D. W. (2010). Qualitative interview design: A practical guide for novice researchers. *The Qualitative Report*, 15(3), 754-760.
- Üztemur, S. S. & Metin, C. (2015). Sosyal bilgiler öğretmenlerinin ölçme değerlendirme alanındaki kavram yanlışları ve öz yeterlik inançlarının incelenmesi. *Anatolian Journal of Educational Leadership and Instruction*, 3(2), 41-67.
- van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *A practical guide to modelling cognitive processes*. London, UK: Academic Press.
- VanLeirsburg, P. & Johns, J. (1991). Assessment literacy: Perceptions of preservice and inservice teachers regarding ethical considerations of standardized testing

procedures. Literacy research report no. 12. Retrieved from <http://files.eric.ed.gov/fulltext/ED341666.pdf>

- Vogel, L. R., Rau, W. C., Baker, P. J., & Ashby, D. E. (2006). Bringing assessment literacy to the local school: A decade of reform initiatives in Illinois. *Journal of Education for Students at Risk (JESPAR)*, 11(1), 39-55. doi:10.1207/s1532767espr1101_3
- Vogt, K. & Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly*, 11, 374-402. doi: 10.1080/15434303.2014.960046
- Vogt, K., Guerin, E., Sahinkarakas, S., Pavlou, P., Tsagari, D., & Afiri, Q. (2008). *Assessment literacy of foreign language teachers in Europe – current trends and future perspectives*. Paper presented at the 5th EALTA Conference, Athens, Greece.
- Volante, L. & Fazio, X. (2007). Exploring teacher candidates' assessment literacy: Implications for teacher education reform and professional development. *Canadian Journal of Education*, 30(3), 749-770.
- Volante, L. & Melahn, C. (2015). Promoting assessment literacy in teachers: Lessons from the Hawai'i school assessment liaison program. *Pacific Educational Research*, 13(1), 103-119.
- Walters, F. S. (2010). Cultivating assessment literacy: Standards evaluation through language-test specification reverse engineering. *Language Assessment Quarterly*, 7(4), 317-342. doi: 10.1080/15434303.2010.516042
- Webb, N. (2002, April). *Assessment literacy in a standards-based education setting*. A paper presented at the annual meeting of the American Educational Research Association, New Orleans, Louisiana.
- White, E. (2009). Are you assessment literate? Some fundamental questions regarding effective classroom-based assessment. *OnCUE Journal*, 3(1), 3-25.
- Williams, J. C. (2015). "Assessing without levels": Preliminary research on assessment literacy in one primary school. *Educational Studies*, 41(3), 341-346. doi:10.1080/03055698.2015.1007926

- Willis, J., Adie, L., & Klenowski, V. (2013). Conceptualising teachers' assessment literacies in an era of curriculum and assessment reform. *The Australian Educational Researcher*, 40, 241-256. doi:10.1007/s13384-013-0089-9
- Witte, R. H. (2010). *Assessment literacy in today's classroom*. Retrieved from <http://www.education.com/reference/article/assessment-literacy-todays-classroom/>
- Xu, Y. & Brown, G. T. L. (2017). University English teacher assessment literacy: A survey-test report from China. *Papers in Language Testing and Assessment*, 6(1), 133-158.
- Yantim, V. & Wongwanich, S. (2014). A study of classroom assessment literacy of primary school teachers. *Procedia – Social and Behavioral Sciences*, 116, 2998-3004.
- Yavuz Kırık, M. (2008). *Yabancı dil olarak İngilizce öğretmenlerinin ölçme değerlendirme bağlamında tutum ve yaklaşımları* (Unpublished doctoral dissertation). İstanbul Üniversitesi, İstanbul, Türkiye.
- Yetkin, C. (2015). *An investigation on ELT teacher candidates' assessment literacy* (Unpublished master thesis). Çağ University, Mersin, Turkey.
- Yin, R. K. (2009). *Case study research: Design and methods* (4th ed.). The United States of America: SAGE Publications, Inc.
- Yıldırım, A. (2012). *İngilizce dil becerilerinin öğretimi ve değerlendirilme durumları arasındaki tutarlığın öğretmen ve öğrenci görüşlerine göre belirlenmesi* (Unpublished master thesis). Ankara Üniversitesi, Ankara, Türkiye.
- Yıldırım, A. & Şimşek, H. (2013). *Sosyal bilimlerde nitel araştırma yöntemleri* (9th ed.). Ankara: Seçkin Yayıncılık.
- Zhang, Z. (1996, April). *Teacher assessment competency: A Rasch Model analysis*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.
- Zhang, Z. & Burry-Stock, J. (1995, November). *A multivariate analysis of teachers' perceived assessment competency as a function of measurement training and*

years of teaching. Paper presented at the annual meeting of the Mid-South Educational Research Association, Biloxi, MS.



APPENDICES

APPENDIX 1. The informed consent form for participating the study

Title: Language Assessment Literacy of Turkish EFL Instructors: A Multiple-case Study

As a participant of this, I have been informed that this study is about language assessment literacy of Turkish EFL instructors.

I know that:

- the first interview focuses on background information in terms of education and teaching approach, assessment and evaluation approach and the factors affecting assessment and evaluation;
- the second interview deals with the sub-components of language assessment literacy: (a) choosing assessment method(s) appropriate for instructional purposes, (b) developing assessments appropriate for instructional purposes, (c) administering assessments, scoring them and interpreting their results, (d) developing a valid grading procedure, (e) using assessment results in decision making, (f) communicating assessment results and recognizing unethical and illegal assessment practices;
- the third interview is going to be carried out in order to analyze the transcriptions of the first and second interviews, think-aloud protocol and focus group discussion and their content analyses;
- the think-aloud protocol is going to be made in order to analyze the transcriptions of the first and second interviews, think-aloud protocol and focus group discussion and their content analyses;
- the focus group discussion is going to be related to the implementation of language assessment literacy;
- the first and second observations are going to be about how we will choose our assessment method(s) and how we will develop our grading system;
- the third observation is going to be on how I am going to administer my exam and about the ethical, legal and professional practices to be followed in administering the exam;
- the fourth observation is going to be made in order to observe how I am going to grade my exam and communicate its results and about the ethical, legal and professional practices to be followed in grading the exam and communicating its results; and
- the document analysis is going to be used for analyzing my quiz, midterm and final exams depending on the document analysis protocol that the researcher has prepared

I know that participating this study is volunteer.

I know that I have the right to withdraw the study in any step.

I have the knowledge about the processes of the study and know what I am supposed to do in the study.

I know that all of my verbal responses will be kept anonymous, used only for this study and stored confidentially. Therefore, I announce that I give my consent.

I verify that I have taken a copy of the consent form.

Participant's Signature

Researcher's Signature

.....

.....

APPENDIX 2. The first semi-structured interview protocol and questions

Date: _____ Nickname: _____

Instructions for the Researcher

1. Introduce yourself.
2. Mention the aim of the study.
3. Provide information about the structure of the interview (recording, taking note and using a nickname).
4. Ask the participant whether s/he has any question to ask.
5. Check the voice recording device.
6. Smile to help the participant feel comfortable.

General Instructions for the Interview

My name is Ahmet Erdost YASTIBAŞ. I am a Ph.D. student in the English language department of the Institute of Educational Sciences of Atatürk University. I am conducting a dissertation study on the topic about which I will inform you.

The information you will give is very important for my study. I thank you for accepting to be interviewed.

My study focuses on language assessment literacy of Turkish EFL instructors. The study aims at finding out the implementation of language assessment literacy, things done during its implementation, the sub-component(s) not implemented and the reason(s) for this.

This phase of the study is related to the participant's background. It focuses on background information of the participant in terms of education and teaching approach, assessment and evaluation approach and the factors affecting the participant's assessment and evaluation practices.

The information you will give is going to be used only for scientific purposes in this study, your participant identity is going to be kept confidential and the findings of this interview are going to be shared with you. The interview lasts almost 60 minutes. You can answer the questions as you want, add or omit when necessary and end the interview whenever you want.

You can send an e-mail to ahmeterdost@gmail.com in order to contact me about the interview.

Before the interview, please read the consent form carefully and sign it.

Questions

Demographic information

1. What would you like to choose as your nickname in order to use during the interview and data analysis?
2. How old are you?
3. From which department did you graduate?
 - a. English language teaching, American culture and literature.
4. Do you have an MA and/or Ph.D. degree or have you been doing an MA and/or Ph.D.? In which field did you have your MA and/or Ph.D. degree or have you been doing your MA and/or Ph.D.?
5. How long have you been working as a teacher?

6. How long have you been working at this university?
7. Did you work at a different school before? If so, what were they?
8. How many hours do you teach every week?
 - a. How many different courses do you teach this term?
9. How many students do you have in your classes?

Education and teaching approach

1. What do education and teaching mean to you?
2. What is your education and teaching approach?
 - a. Like democratic, teacher-centered, or student-centered.
3. What is/are the factor(s) affecting your understanding of education and teaching?
 - a. Like a book read, memory, or a person taken as a model.

Assessment and evaluation approach

1. What do assessment and evaluation mean to you? Explain.
 - a. Why?
2. What kind of associations do assessment and evaluation create? Why?
 - a. Like tiring, exciting, or exploratory.
3. How do you feel when you assess and evaluate? What do you think causes this/these feeling(s)?
 - a. Like anger, happiness, or control.
4. What does assessment literacy mean to you?
 - a. Why?
5. Do you think assessment and evaluation have an effect on your education and teaching approach and practices? If so, what is/are this/these effect(s)?
 - a. (To exemplify) How do you feel and what do you do when your exam results are very low and high?

Factors affecting assessment and evaluation

1. **A.** How were you assessed and evaluated when you were a student?
 - a. Like multiple-choice, writing, or project.
 - b. Primary, elementary, high schools and university periods (MA and/or Ph.D. periods can also be asked according to the answer of the fourth question in the demographic information part)
- B.** What do you think about these assessment methods and tools?
- C.** Did this/these assessment method(s) affect you when you were a student? If so, how?
 2. **A.** Have you taken pre-service assessment and evaluation training? If so, what did you study?
 - a. Like assessment and evaluation in education, statistics, or language assessment and evaluation course.
 - B.** Do you think pre-service assessment training has an effect on your present assessment and evaluation practices?
 - a. If so, what is/are its effect(s)?
 - b. If not, why do you think it does not affect your practices?

3. A. Have you taken any in-service assessment training? If so, what kind of training have you taken?
- B. Do you think pre-service assessment training has an effect on your present assessment and evaluation practices?
 - a. If so, what is/are its effect(s)?
 - b. If not, why do you think it does not affect your practices?
4. (IF ANY TRAINING HAS NOT BEEN TAKEN + EVEN IF PRE-SERVICE AND/OR IN-SERVICE TRAINING HAS BEEN TAKEN) How have you improved yourself in assessment and evaluation?
 - a. What kind of sources have you read?
 - b. Have you interacted with your peers to improve yourself?
5. A. Is there any difference between your assessment approach at the beginning of your career and your approach now?
 - a. If so, what is/are its effect(s)?
- B. What do you think causes this/these difference(s)?
- C. Is/Are the difference(s) positive or negative?
 - a. Explain the positive and/or negative change(s).
6. Why do you assess and evaluate your students?
 - a. Like giving grades or developing instructional materials.
7. What do you pay attention to in assessing and evaluating your students?
 - a. Why?
8. A. Do you encounter any difficulty in assessing and evaluating your students? If so, what is/are it/they?
 - a. Why?
- B. How do you overcome this/these difficulty(ies)?

Closure

The interview is ended. Do you want to add to and omit from your explanation?

I thank you for participating the interview voluntarily, sparing your valuable time and answering my questions. As I have mentioned before, if you want to contact me about the interview, you can send an e-mail to ahmeterdost@gmail.com.

APPENDIX 3. The second semi-structured interview protocol and questions

Date: _____ Nickname: _____

Instructions for the Researcher

1. Introduce yourself.
2. Mention the aim of the study.
3. Provide information about the structure of the interview (recording, taking note and using a nickname).
4. Ask the participant whether s/he has any question to ask.
5. Check the voice recording device.
6. Smile to help the participant feel comfortable.

General Instructions for the Interview

My name is Ahmet Erdost YASTIBAŞ. I am a Ph.D. student in the English language department of the Institute of Educational Sciences of Atatürk University. I am going to inform you about the second part of my dissertation study.

The information you will give is very important for my study. I thank you for accepting to be interviewed.

The second interview deals with the sub-components of language assessment literacy. The sub-components involve (a) choosing assessment method(s) appropriate for instructional purposes, (b) developing assessments appropriate for instructional purposes, (c) administering assessments, scoring them and interpreting their results, (d) developing a valid grading procedure, (e) using assessment results in decision making, (f) communicating assessment results and recognizing unethical and illegal assessment practices.

The information you will give is going to be used only for scientific purposes in this study, your participant identity is going to be kept confidential and the findings of this interview are going to be shared with you. The interview lasts almost 75 minutes. You can answer the questions as you want, add or omit when necessary and end the interview whenever you want.

You can send an e-mail to ahmeterdost@gmail.com in order to contact me about the interview.

Before the interview, please read the consent form carefully and sign it.

Questions

The First Sub-component: Choosing Assessment Method(s) According to Instructional Purposes

1. What does measurement error mean to you? Could you please explain it by giving an example?
 - a. Like measurement error resulting from the assessment tool, student and teacher.
 - b. Like systematic, constant and random error.
2. What does validity mean to you?
 - a. Like face, construct, or content validity.

3. What do valid and invalid assessment data mean to you? Do you think valid and invalid assessment data affect instructional activities and decisions? If so, how do they affect?
4. What sort of assessment method(s) have you used since you started teaching?
 - a. Selected-response
 - b. Constructed-response
 - c. Performance assessment
 - d. Informal personal communication
5. What sort of assessment method(s) have you decided to use for your present courses?
6. What did you pay attention to in choosing your assessment method(s)?
 - a. (ASK IF NOT MENTIONED) Did you pay attention to measurement error and validity?
 - b. Do you make any research about the method(s) before choosing it/them? If so, what kind of sources do you use?
7. For what purpose(s) do you use the chosen assessment method(s)?
 - a. Like measuring information, analyzing, or making a synthesis.
8. What do you think about the strength(s) and weakness(es) of the assessment method(s) you have chosen?

The Second Sub-component: Developing Assessment(s) Appropriate for Instructional Purposes

1. How do you prepare your exam(s) relevant to the assessment method(s) you have chosen?
 - a. The things paid attention to.
 - b. Detailed information about the preparation process.
2. What type(s) of questions do you use in your exams? Why do you use it/them?
3. How do you make your exams valid?

The Third Sub-component: Administering Exams, Scoring Them and Interpreting Their Results

1. What does reliability mean to you?
2. How do you make your exams reliable?
3. How do you administer your exams?
 - a. What do you pay attention to in administering your exams?
 - b. How do you deal with any problem you encounter in administering your exams?
4. How do you score your exams?
 - a. What do you pay attention to in scoring your exams?
5. How do you grade your exams?
 - a. What do you pay attention to in grading your exams?
6. What does the consistency of interpreting exams mean to you?
 - a. Why?
 - b. (If the concept is known) How do you make your interpretations consistent?

- c. (If the concept is known) What do you do in case of inconsistency in interpreting the results?
- 7. How do you interpret formal and informal students' assessment results?
 - a. Do you use the results to improve your assessment tool? How?
 - b. Do you use the results to find out and improve your students' weaknesses? How?
- 8. Do you take into account the effect(s) of your exams on your students?
 - a. If yes, how?
 - b. If no, why?
- 9. How do you keep your exams and exams' results confidential?
- 10. What is your attitude toward exam complaint?
 - a. How do your students make exam complaints?

The Fourth Sub-component: Using Assessment Results in Making Decisions about Student, Education, Curriculum and School

- 1. How should assessment and evaluation data be evaluated according to you?
 - a. What does correct and wrong evaluation mean to you? Please, give examples.
- 2. Do you use assessment and evaluation data to improve instructional plans to improve your students' learning?
 - a. If yes, how?
 - b. If no, why?
- 3. Do you make any change in your instruction and curriculum?
 - a. If yes, how?
 - b. If no, why?

The Fifth Sub-component: Developing a Valid Grading Procedure for Assessing and Evaluating Students

- 1. What is/are the grading system(s) in your courses?
 - a. Why do you use this/these system(s)?
 - b. How have you decided its/their components?
- 2. If you have not developed any grading system for your courses, what do you think about the grading system(s) given to you?
 - a. Did you take part in its/their development process?
 - i. If so, how was/were the grading system(s) developed?
 - b. Do you use the available grading system without making any change?
 - i. If not, why? Do you make any change? Please, explain.
- 3. What do you think about the validity of the grades you give to your students by using this/these grading system(s)? How do you explain the grades' validity to your students?
- 4. Why do you use the grades?

The Sixth Sub-component: Communicating Assessment Results to Students and Other Stakeholders

- 1. How do you announce your exam results to your students and administrators?

2. What do the results mean to your students and administrators according to you?
 - a. How should assessment results be evaluated?
 - i. Do you think the chosen assessment method(s) has/have an effect on evaluating the results?
 - b. Is there any limitation in evaluating the results? If so, what is/are it/they?
 - c. What do you think about the reflection of evaluating the results?
 - d. How do you think the misinterpretation and misevaluation of assessment results can be avoided?
 - i. Do you inform your students and administrators about how to evaluate assessment data?
3. How do you deal with any possible measurement error in communicating your assessment results?

The Seventh Sub-component: Recognizing any Unethical and Illegal Assessment and Evaluation Practices

1. What do you think about the ethical, legal and professional behaviors a teacher should follow in assessing and evaluating his students?
2. Do you encounter any problem in terms of your answer to the first question? What is/are the problem(s)? How do you deal with this/these problem(s)?

Closure

The interview is ended. Do you want to add to and omit from your explanation?

I thank you for participating the interview voluntarily, sparing your valuable time and answering my questions. As I have mentioned before, if you want to contact me about the interview, you can send an e-mail to ahmeterdost@gmail.com.

APPENDIX 4. The third semi-structured interview protocol and questions

Date: _____ Nickname: _____

Instructions for the Researcher

1. Introduce yourself.
2. Mention the aim of the study.
3. Provide information about the structure of the interview (recording, taking note and using a nickname).
4. Ask the participant whether s/he has any question to ask.
5. Check the voice recording device.
6. Smile to help the participant feel comfortable.

General Instructions for the Interview

My name is Ahmet Erdost YASTIBAŞ. I am a Ph.D. student in the English language department of the Institute of Educational Sciences of Atatürk University. I am going to inform you about this part of my dissertation study.

The information you will give is very important for my study. I thank you for accepting to be interviewed.

The third interview is about what you think about the transcriptions of the think-aloud protocol, first and second interviews and focus group discussion and their content analyses.

The information you will give is going to be used only for scientific purposes in this study, your participant identity is going to be kept confidential and the findings of this interview are going to be shared with you. The interview lasts almost 30 minutes. You can answer the questions as you want, add or omit when necessary and end the interview whenever you want.

You can send an e-mail to ahmeterdost@gmail.com in order to contact me about the interview.

Before the interview, please read the consent form carefully and sign it.

Questions

1. Is there anything that you want to add to and/or omit from your explanations in the transcription of the first interview? If so, please explain what you want to add and/or omit by mention your reason(s) for it/them.
2. Do you agree with the findings of the first interview? If not, what and why do you disagree?
3. Is there anything that you want to add to and/or omit from your explanations in the transcription of the second interview? If so, please explain what you want to add and/or omit by mention your reason(s) for it/them.
4. Do you agree with the findings of the second interview? If not, what and why do you disagree?
5. Is there anything that you want to add to and/or omit from your explanations in the transcription of the think-aloud protocol? If so, please explain what you want to add and/or omit by mention your reason(s) for it/them.
6. Do you agree with the findings of the think-aloud protocol? If not, what and why do you disagree?

7. Is there anything that you want to add to and/or omit from your explanations in the transcription of the focus group discussion? If so, please explain what you want to add and/or omit by mention your reason(s) for it/them.
8. Do you agree with the findings of the parts related to you in the focus group discussion? If not, what and why do you disagree?

Closure

The interview is ended. Do you want to add to and omit from your explanation?

I thank you for participating the interview voluntarily, sparing your valuable time and answering my questions. As I have mentioned before, if you want to contact me about the interview, you can send an e-mail to ahmeterdost@gmail.com.



APPENDIX 5. Think-aloud protocol

Date: _____ **Participants:** _____

General Instructions

Think-aloud protocol is talking thoughts aloud while an action is performed. During this process both the action is performed and thoughts are voiced.

I want you to share your thoughts with me while preparing your midterm exam. During this process, any question is not going to be asked to you. How you will prepare your questions is going to be observed and what you will talk aloud about your thoughts in writing questions is going to be recorded.

The information you will give is going to be used only for scientific purposes in this study, your participant identity is going to be kept confidential and the findings of the think-aloud protocol are going to be shared with you. You can end the think-aloud protocol whenever you want.

You can send an e-mail to ahmeterdost@gmail.com in order to contact me about the interview.

Before the think-aloud protocol, please read the consent form carefully and sign it.

Think-aloud Protocol Instructions

1. Be ready for preparing your exam.
2. Talk what you are going to do and what you are going to think aloud.
3. You are not going to be disturbed while preparing your exam.
4. The main goal of the think-aloud protocol is not to judge how you prepare your exams, but to reveal the mental processes you use in preparing your questions.

APPENDIX 6. The focus group discussion protocol and questions

Date: _____ Nickname: _____

Instructions for the Researcher

1. Introduce yourself.
2. Mention the aim of the study.
3. Provide information about the structure of the interview (recording, taking note and using a nickname).
4. Ask the participant whether s/he has any question to ask.
5. Check the voice recording device.
6. Smile to help the participant feel comfortable.

General Instructions for the Interview

My name is Ahmet Erdost YASTIBAŞ. I am a Ph.D. student in the English language department of the Institute of Educational Sciences of Atatürk University. I am conducting a dissertation study on the topic about which I will inform you.

The information you will give is very important for my study. I thank you for accepting to be interviewed.

My study focuses on language assessment literacy of Turkish EFL instructors. The study aims at finding out the implementation of language assessment literacy, things done during its implementation, the sub-component(s) not implemented and the reason(s) for this.

This phase of the study is related to the participant's background. It focuses on background information of the participant in terms of education and teaching approach, assessment and evaluation approach and the factors affecting the participant's assessment and evaluation practices.

The information you will give is going to be used only for scientific purposes in this study, your participant identity is going to be kept confidential and the findings of this interview are going to be shared with you. The focus group discussion lasts almost 60 minutes. You can answer the questions as you want, add or omit when necessary and end the focus group discussion whenever you want.

You can send an e-mail to ahmeterdost@gmail.com in order to contact me about the interview.

Before the focus group discussion, please read the consent form carefully and sign it.

Questions

1. How do you think education and teaching approach affects assessment and evaluation approach and practices?
2. How have assessment and evaluation courses in the pre-service, in-service and/or pedagogical training affected your assessment and evaluation practices?
3. How do you think peer interaction, working experience and experience as a student affect assessment and evaluation approach and practices?
4. How do assessment and evaluation affect a teacher's instruction?
5. How should an/- assessment method(s) be chosen depending on instructional decisions?
6. How should an assessment be developed depending on instructional decisions?

7. How should an assessment be administered?
8. How should an assessment be scored?
9. How should assessment results be interpreted?
10. How should assessment results be used in making decisions about students, curriculum, school and instruction?
11. How should a valid grading procedure be developed?
12. How should assessment results be communicated to students and administrators?
13. What is/are unethical and illegal practice(s) in assessment and evaluation? How should it/they be dealt with?
14. What does language assessment literacy mean to you?
15. How should a language assessment literacy course be designed?

Closure

The interview is ended. Do you want to add to and omit from your explanation?

I thank you for participating the focus group discussion voluntarily, sparing your valuable time and answering my questions. As I have mentioned before, if you want to contact me about the interview, you can send an e-mail to ahmeterdost@gmail.com.

APPENDIX 7. The first and second non-participant observation protocols

The Aim of the Study	To find out how Turkish EFL instructors implement their language assessment literacy in their English classes.	
Research Questions	<ol style="list-style-type: none"> 1. How do Turkish EFL instructors implement the sub-components of language assessment literacy in their English classes? 2. Which factors affects Turkish EFL instructors' implementation of LAL in the class? 3. What are the effects of LAL on Turkish EFL instructors? 4. Do Turkish EFL instructors encounter any difficulty while implementing LAL? If so, what are they? How do they overcome them? 5. Do Turkish EFL instructors implement all sub-components of language assessment literacy? If not, which sub-component is it or which sub-components are they? What causes it or them? 	
General Information about the Observation	Explain when and where the observation is going to be made with whom.	Place:
		Date:
		The Participants:
		The Observer:
Information about the Observation Place	Give information about the number of the participants in the meeting room and about the materials to used in the meeting.	
Phenomena to Be Observed	Phenomena to be observed are going to be written clearly here.	The phenomena are choosing an assessment method and developing a grading system.
The Sub-dimensions of Phenomena to Be Observed	The sub-dimensions to be paid attention during the observation are going to mentioned here. They are going to be focused on in the observation.	The Sub-dimensions of the First and Second Phenomenon <ol style="list-style-type: none"> 1. Things to be done at the beginning of the meeting. 2. How the meeting will be directed. 3. What the participants will do during the meeting. 4. How the decisions are going to be made.

APPENDIX 8. The third non-participant observation protocol

The Aim of the Study	To find out how Turkish EFL instructors implement their language assessment literacy in their English classes.		
Research Questions	<ol style="list-style-type: none"> 1. How do Turkish EFL instructors implement the sub-components of language assessment literacy in their English classes? 2. Which factors affects Turkish EFL instructors' implementation of LAL in the class? 3. What are the effects of LAL on Turkish EFL instructors? 4. Do Turkish EFL instructors encounter any difficulty while implementing LAL? If so, what are they? How do they overcome them? 5. Do Turkish EFL instructors implement all sub-components of language assessment literacy? If not, which sub-component is it or which sub-components are they? What causes it or them? 		
General Information about the Observation	Explain when and where the observation is going to be made with whom.	Place:	
		Date:	
		The Participant:	
		The Observer:	
Information about the Observation Place	Give information about the number of the students, of the desks and of the proctors in the exam room and about the materials to use in the exam.		
Phenomena to Be Observed	Phenomena to be observed are going to be written clearly here.	The phenomena to be observed are how the exam will be administered and what kind of ethical, legal and professional assessment practices are going to be followed.	
The Sub-dimensions of Phenomena to Be Observed	The sub-dimensions to be paid attention during the observation are going to mentioned here. They are going to be focused on in the observation.	The Sub-dimensions of the First Phenomenon <ol style="list-style-type: none"> 1. Things done before and when the exam starts. 2. Things done during and after the exam. 3. The interaction and communication between the students and the participant during the whole process. 4. The interaction and communication between the proctors during the whole process. 5. The behaviors of the participant during the whole process. 	The Sub-dimensions of the First Phenomenon <ol style="list-style-type: none"> 1. The ethical, professional and legal behaviors the participant has mentioned in the first and second interviews. 2. Any problem encountered in administering the exam. 3. The things occurring during any problem: <ul style="list-style-type: none"> - The attitude of the participant. - The reaction of the participant.

APPENDIX 9. The fourth non-participant observation protocol

The Aim of the Study	To find out how Turkish EFL instructors implement their language assessment literacy in their English classes.			
Research Questions	<ol style="list-style-type: none"> 1. How do Turkish EFL instructors implement the sub-components of language assessment literacy in their English classes? 2. Which factors affects Turkish EFL instructors' implementation of LAL in the class? 3. What are the effects of LAL on Turkish EFL instructors? 4. Do Turkish EFL instructors encounter any difficulty while implementing LAL? If so, what are they? How do they overcome them? 5. Do Turkish EFL instructors implement all sub-components of language assessment literacy? If not, which sub-component is it or which sub-components are they? What causes it or them? 			
General Information about the Observation	Explain when and where the observation is going to be made with whom.	Place:		
		Date:		
		The participant:		
		The Observer:		
Information about the Observation Place	Describe the office in which the participant is going to grade his/her exams. Give information about the system used for announcing students' grades.			
Phenomena to Be Observed	Phenomena to be observed are going to be written clearly here.	How the exam is going to be graded, how its results are going to be announced and what kind of ethical, legal and professional assessment practices are going to be followed are going to be observed.		
The Sub-dimensions of Phenomena to Be Observed	The sub-dimensions to be paid attention during the observation are going to be mentioned here. They are going to be focused on in the observation.	The Sub-dimensions of the First Phenomenon <ol style="list-style-type: none"> 1. Things done at the beginning of grading the exam. 2. Things done in grading the exam. 3. Things done at the end of grading the exam. 	The Sub-dimensions of the Second Phenomenon <ol style="list-style-type: none"> 1. Things done at the beginning of communicating the results. 2. Things done in communicating the results. 3. Things done at the end of communicating the results. 	The Sub-dimensions of the Third Phenomenon <ol style="list-style-type: none"> 1. The ethical, professional and legal behaviors the participant has mentioned in the first and second interviews. 2. Any problem encountered in grading the exam and communicating its results 3. The thing(s) occurring during any problem. 4. The attitude and reaction of the participant.

APPENDIX 10. The document analysis protocol

In addition to the discussion above about the key concepts of testing and assessment, Brown (2004) mentioned that there are five basic principles of language assessment: practicality, reliability, validity, authenticity and washback. He told that these five principles can be applied to the evaluation of language assessment by asking the following six questions (p. 30):

1. Are the test procedures practical?
2. Is the test reliable?
3. Does the procedure demonstrate construct validity?
4. Is the procedure face valid and biased for best?
5. Are the test tasks as authentic as possible?
6. Does the test offer beneficial washback to the learner?

For the first question, Brown (2004) proposed a practicality checklist which was composed of seven questions. They were:

1. “Are the administrative details clearly established before the test;
2. Can students complete the test reasonably within the set time frame;
3. Can the test be administered smoothly without practical glitches;
4. Are all materials and equipment ready;
5. Is the cost of the test within budgeted limits;
6. Is the scoring/ evaluation system feasible within the teacher’s time frame; and
7. Are the methods for reporting results determined in advance (Brown, 2004, p. 31)?

For the second question, Brown (2004) mentioned that a teacher could ensure that “all students receive the same quality of input, whether written or auditory” (p.31). Therefore, the teacher should pay maximum attention to the quality of the copied papers, classroom conditions, the audibility of the listening material and objective scoring procedures (Brown, 2004). For rater reliability, especially for intra-rater reliability, Brown (2004) stressed out that the teacher should not allow his/her stamina and concentration to reduce. In addition, transparency is considered as an important part of reliability (Rogier, 2014). When the assessment is made transparent, his students know what they have to learn and how their learning is assessed (Rogier, 2014).

For the third question, the teacher had better be sure that his/her classroom objectives are identified, appropriately framed and shown in the test-specifications table (Brown, 2004). If the teacher is sure that he/she had done these, it means his/her exam is content valid.

In accordance with the fourth question, Brown (2004) stated that if a teacher gives enough time to his/her students to study for his/her exam, he/she can avoid any bias that might affect his/her students negatively. Besides, Brown mentioned that if an exam meets the following needs below, it becomes face valid. The criteria are:

1. Instructions are clear for students to understand;
2. The exam is structured and organized logically;
3. The exam is appropriate to students’ level;
4. The exam does not include surprises for students; and
5. The exam has appropriate timing for students (Brown, 2004).

For the fifth question, Brown (2004) mentioned that the teacher should be sure that he/she uses the language as natural as possible in his/her exam, that he/she contextualizes items in his/her exam, that he/she selects interesting, enjoyable and

humorous topics, that he/she provides some thematic organization and that he/she uses tasks that are similar to real-world tasks.

For the last question, the teacher can produce a positive washback if his/her exam is content valid (Brown, 2004). The other thing that he/she can do is to give his/her students enough time to study for the exam (Brown, 2004). To provide a positive washback effect, an assessment must be aligned with the goal and objectives of a course (Rogier, 2014). Besides, the assessment must reflect the teaching activities (Rogier, 2014).



CURRICULUM VITAE

Personal Information

Name – Surname: Ahmet Erdost YASTIBAŞ

Place and Date of Birth: Mut / 01.01.1987

Education

Primary School: Mareşal Fevzi Çakmak Primary School – 1994 – 2001

High School: Mut Anatolian High School – 2001 – 2005

University: Mersin University – Faculty of Education – English Language Teaching Department – 2005 – 2009

Master: Çağ University – Graduate School of Social Sciences – English Language Teaching Department – 2011 – 2013

Contact Information

Email Address: ahmeterdost@gmail.com