DEVELOPING AND VALIDATING LANGUAGE ASSESSMENT KNOWLEDGE SCALE (LAKS) AND EXPLORING THE ASSESSMENT KNOWLEDGE OF EFL TEACHERS

Doktora Tezi

Elçin ÖLMEZER-ÖZTÜRK

Eskişehir 2018

DEVELOPING AND VALIDATING LANGUAGE ASSESSMENT KNOWLEDGE SCALE (LAKS) AND EXPLORING THE ASSESSMENT KNOWLEDGE OF EFL TEACHERS

Elçin ÖLMEZER-ÖZTÜRK

PhD DISSERTATION
Programme in English Language Teaching
Supervisor: Prof. Dr. Belgin AYDIN

Eskişehir
Anadolu University
Graduate School of Educational Sciences
July 2018

Bu tez çalışması BAP Komisyonunca kabul edilen 1706E367 no.lu proje kapsamında desteklenmiştir.

T.C.
ANADOLU ÜNİVERSİTESİ
Eğitim Bilimleri Enstitüsü Müdürlüğü

JÜRİ VE ENSTİTÜ ONAYI

Elçin ÖLMEZER ÖZTÜRK'ün "Developing And Validating Language Assessment Knowledge Scale (LAKS) And Exploring The Assessment Knowledge of EFL Teachers" başlıklı tezi 02.07.2018 tarihinde aşağıdaki jüri tarafından değerlendirilerek "Anadolu Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliği"nin ilgili maddeleri uyarınca Yabancı Diller Eğitimi Anabilim Dalı İngilizce Öğretmenliği Doktora Programında, Doktora tezi olarak kabul edilmiştir.

	Unvanı-Adı Soyadı	İmza
Üye (Tez Danışmanı)	: Prof. Dr. Belgin AYDIN	
Üye	: Prof. Dr. Dinçay KÖKSAL	
Üye	:Prof. Dr. Gülsev PAKKAN	
Üye	: Doç. Dr. Özgür YILDIRIM	
Üye	: Doç. Dr. Murat AKYILDIZ	

Prof.Dr. Handan DEVECİ
Anadolu Üniversitesi
Eğitim Bilimleri Enstitüsü Müdürü

ABSTRACT

DEVELOPING AND VALIDATING LANGUAGE ASSESSMENT KNOWLEDGE SCALE (LAKS) AND EXPLORING THE ASSESSMENT KNOWLEDGE OF EFL TEACHERS

Elçin ÖLMEZER-ÖZTÜRK

Department of Foreign Language Education, Programme in English Language Teaching

Anadolu University, Graduate School of Educational Sciences, July 2018

Supervisor: Prof. Dr. Belgin AYDIN

The aim of this study is two-fold: to develop and validate Language Assessment Knowledge Scale (LAKS) as an instrument to measure language assessment knowledge (LAK) of EFL teachers and to provide a general picture regarding LAK level of these teachers working in Turkish higher education context. After a thorough validation process, LAKS with 60 items and four constructs (assessing reading, assessing listening, assessing writing, and assessing speaking) was answered by 542 EFL teachers working in higher education context. As for the qualitative phase, 11 teachers provided detailed answers to open-ended questions that were asked to get in-depth data regarding teachers' opinions on language assessment knowledge. The statistical findings regarding the validity and reliability of the scale revealed that LAKS had a perfect model-data fit and Cronbach Alpha coefficients were high. In terms of LAK level of the teachers, the participants got, on average, 25 out of 60, significantly lower than half of the total score. It was also found that the teachers were the most knowledgeable in assessing reading whereas they had the lowest score in assessing listening. Besides, except for being a testing office member or not, no significant impact of demographic features was found on LAK level of the participants. On the other hand, the qualitative findings showed that education in pre-service and in-service levels were insufficient, and teachers needed trainings on assessing each skill specifically. Finally, the present study offers several suggestions both for future studies and for policy makers to improve EFL teachers' language assessment literacy.

Keywords: Language assessment literacy, Language assessment knowledge, Language testing and assessment, EFL teachers.

ÖZET

DİLDE ÖLÇME DEĞERLENDİRME BİLGİSİ ÖLÇEĞİNİN GELİŞTİRİLMESİ VE İNGİLİZCE ÖĞRETMENLERİNİN DİLDE ÖLÇME DEĞERLENDİRME BİLGİLERİNİN İNCELENMESİ

Elçin ÖLMEZER-ÖZTÜRK
Yabancı Diller Eğitimi Anabilim Dalı, İngilizce Öğretmenliği Programı
Anadolu Üniversitesi, Eğitim Bilimleri Enstitüsü, Temmuz 2018
Danışman: Prof. Dr. Belgin AYDIN

Bu çalışma, öğretmenlerin dilde ölçme değerlendirme bilgilerini ölçmek için Dilde Ölçme Değerlendirme Bilgisi Ölçeği'ni geliştirmeyi ve Türkiye'de yükseköğretim bağlamında çalışan İngilizce öğretmenlerinin dilde ölçme değerlendirme bilgisi seviyelerini ortaya koymayı amaçlamaktadır. Kapsamlı bir geçerlilik çalışması sürecinden sonra 60 maddeli ve 4 boyutlu bu ölçek, yükseköğretimde çalışan 542 öğretmen tarafından cevaplanmıştır. Çalışmanın nitel veri toplama evresinde ise, öğretmenlerin dilde ölçme değerlendirme ile alakalı ayrıntılı görüşlerini elde etmek için 11 öğretmen önceden hazırlanmış açık uçlu sorulara detaylı cevaplar vermişlerdir. Ölçeğin geçerliliği ve güvenirliği ile ilgili istatistiksel bulgular ölçeğin mükemmel model veri uyumuna sahip olduğunu ve güvenirliğin yüksek olduğunu ortaya koymuştur. Öğretmenlerin dilde ölçme değerlendirme bilgi seviyeleri ile alakalı ise, katılımcıların ölçeğin genelinden 60 üzerinden 25 aldığı ve bu ortalamanın toplam puanın yarısından anlamlı derecede düşük olduğu sonucuna varılmıştır. Ölçme değerlendirme ofisi çalışanı olup olmama haricinde hiçbir demografik değişkenin öğretmenlerin dilde ölçme değerlendirme bilgisi üzerinde etkisinin olmadığı da çalışmanın istatistiki bulguları arasındadır. Buna ek olarak, çalışmanın nitel verileri hizmet öncesi ve sonrasında verilen ölçme değerlendirme eğitiminin ciddi anlamda yetersiz olduğunu ve öğretmenlerin her bir becerinin ölçülmesi üzerine hizmet-içi eğitimlere ihtiyaç duyduğunu ortaya koymuştur.

Anahtar Sözcükler: Dilde ölçme değerlendirme okuryazarlığı, Dilde ölçme değerlendirme bilgisi, Dilde ölçme ve değerlendirme, İngilizce öğretmenleri.

ACKNOWLEDGEMENTS

First of all, I would like to thank my supervisor, Prof. Dr. Belgin AYDIN, who is one of the most precious people I have ever met. She means a lot to me. Indeed, there exists no vocabulary item in the dictionary to be used while describing her, but still I will try to describe her. In this tedious, informative, and demanding process, she helped me a lot with her smiling face, motivation, willingness, hope, and extensive knowledge. By means of her help and continuous guidance, I could write this PhD dissertation. Whatever I write here to describe her as a supervisor and a person, the sentences seem incomplete to me… I feel myself really lucky to have met her. Last but not the least, she is like the *people in white* in Turkish movies, who appear all of a sudden through the smokes, open their hands, and assist people. She has changed my whole life in such a positive way that I had never imagined.

Besides, I am grateful to the committee members who are Assoc. Prof. Dr. Murat AKYILDIZ and Assoc. Prof. Dr. Özgür YILDIRIM. With their great knowledge, expertise in the field, constructive feedback, and positive attitudes, this thesis got better. They were also very helpful and gave me a hand whenever I needed. Thus, they became more than committee members to me… I am also grateful to the other committee members, Prof. Dr. Dinçay KÖKSAL and Prof. Dr. Gülsev PAKKAN for their acceptance to be in my jury and their support.

Very special thanks go to Dr. Murat Doğan ŞAHİN for his great help in the statistical analyses of my thesis. He was there all the time whenever I needed help and had a question.

Additionally, I would like to thank TÜBİTAK for the scholarship they provided me from the beginning of my PhD, and BAP for the fund they provided us for the materials and equipments needed for this study.

Other people for whom I am having great difficulty finding the appropriate words to describe them and their support to me are my family: Hatice ÖLMEZER, Erdoğan ÖLMEZER, Gökhan ÖZTÜRK and İpek ÖZTÜRK. My mother, Hatice ÖLMEZER, and my father, Erdoğan ÖLMEZER, were always there to help me with their motivating words, smiling faces and supports. They cooked for us, took care of my daughter while I was studying, etc. In short, they did everything more than parents should do for their

child. They believed in me wholeheartedly in each and every second of my life. Thanks to them, I was able to follow my dreams…

My husband, Dr. Gökhan ÖZTÜRK, was with me all the time. Whenever I needed him, he was there just looking at me and waiting for helping me. I am grateful to him for his unconditional love, support, patience, smiling face, motivation, understanding, empathy, shining eyes, taking care of our daughter, and expertise in guiding me in this process, and for everything he did just to please me… He is more than a friend, a husband, and a colleague…

The last thanks go to my lovely daughter, İpek ÖZTÜRK. When I started my PhD programme, she was just one year old. I tried a lot for my studies not to affect her negatively, because she was just a baby. I did my best to be with her and have fun together. This process was full of happiness, nervousness, sadness, frustration and tiredness for me as a person trying to have a balance between being a mother and an academician… When she complained about my studying too much, I was broken into pieces. She always asked "Have you passed your exam?" and when I said yes, she got happy thinking that it was the last exam and I would not study again. When she saw me studying, her disspointment started again. Sorry, honey… I have missed many moments that could have been spent with you, but all of these efforts are for our family and for you, don't forget this… Now, sweetie, this PhD process is over, and we are here at the end of this PhD journey… and I am writing my acknowledgement… Thank God… I love you so much ömrüm, nefesim, en kıymetlim…

Elçin ÖLMEZER-ÖZTÜRK

Eskişehir 2018

02.07.2018

STATEMENT OF COMPLIANCE WITH ETHICAL PRINCIPLES AND RULES

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

Elçin ÖLMEZER-ÖZTÜRK

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

Page

LIST OF FIGURES

LIST OF ABBREVIATIONS

AL	: Assessment Literacy
EFL	: English as a Foreign Langauge
ELT	: English Language Teaching
LAK	: Language Assessment Knowledge
LAKS	: Language Assessment Knowledge Scale
LAL	: Language Assessment Literacy
LTA	: Language Testing and Assessment

1. INTRODUCTION

1.1. Background to the Study

The role of assessment in teaching and learning process is undeniable. Assessment is regarded like an engine which is responsible for initiating learning (White, 2009). Recently, there has been a shift for the use of the term assessment; and testing as a term, which had been popular till a few years ago, has been replaced by the term assessment (Inbar- Lourie, 2008). Brown (2003) drew attention to a point which is the common belief that testing and assessment have the same meaning, but it is indeed not. He made a distinction between these two terms. Tests are related to administrative issues and learners know that they are going to be evaluated; however, assessment is "an ongoing process that encompasses a much wider domain" (p. 15). Similarly, Clapham and Corson (1997) pointed out that testing and assessment are different terms, former designed for large number of people, and latter referring to a kind of evaluation whose primary concern is not to get scores; rather, it is usually carried out individually in order to detect learners' problems. When it comes to the process of teaching and learning a second or foreign language, assessment becomes more specific and to the point, and language assessment comes to the ground. Purpura (2016, p. 191) defined language assessment as "a broad term referring to a systematic procedure for eliciting test and nontest for the purpose of making inferences or claims about certain language-related characteristics of an individual". How language testing and assessment are perceived has been subjected to many changes till now, and the differences in the understanding of language testing and assessment have been closely related to the changes in theories in language education.

1.2. Approaches to Language Testing and Assessment

Language testing and assessment have undergone many changes throughout the years. The term testing was popular and the preferred one in the past; thus, the history of language testing has witnessed many shifts. According to Heaton (1990), there are four major approaches in language testing which are the essay-translation approach, the structuralist approach, the integrative approach and the communicative approach. These shifts in language testing are in parallel with the teaching methodology.

Essay translation-approach is also called as pre-scientific stage of language testing. In this approach, grammar-translation is dominant, and the most important thing is the subjective evaluation of the teacher, and the teacher is not expected to have special skills

in testing. Tests are often full of essays, translation, and grammatical analysis. The language used in tests is full of literary and cultural items.

In the structuralist approach, the prevalent understanding is that learning is the acquisition of a set of habits; thus, there is a special emphasis on contrastive analysis. No context is necessary to test language skills, and language elements such as grammar, phonology and lexicon are tested separately. The language skills that are reading, writing, listening and speaking are also tested separately; that is, one skill should be tested at a time. Thus, the tests that are used a lot are analytical and discrete tests. The structuralist approach has a psychometric basis, with a focus on the tests that are appropriate for statistical analyses by giving importance to the need for reliability and validity. Hence, multiple-choice items that are suitable for statistical analyses are popular in this approach. This approach is in contrast with essay-translation approach which is regarded as too subjective and unreliable.

In the integrative approach, meaning and context have gained importance; so, the idea is that skills should not be tested separately, but rather, two or more skills could be tested at the same time. To what extent learners are able to use many skills at the same time is tested in this approach. Integrative tests "are concerned with a global view of proficiency- an underlying language competence or grammar of expectancy, which is argued every learner possesses" (Heaton, 1990, p. 16), and the reason behind learning a language is not important. This view is broader than the previous ones, covering essay writing, translation, and interviews. It has come out in contrast with the logic of discrete test type. With the prevalence of integrative approach, cloze testing and dictation have become popular. The logic behind cloze test is Gestalt theory. The aim here is to be able to make use of all the clues given in context by decoding interrupted messages. Another popular form is dictation, which is previously used for only listening abilities. However, it is understood that dictation involves much more than this, and it requires "auditory discrimination, the auditory memory span, spelling, the recognition of sound segments, a familiarity with the grammatical and lexical patterning of the language, and overall textual comprehension" (Heaton, 1990, p. 17). Despite all the forms of integrative testing, as Lewkowicz (1997) stressed, there has occurred a mismatch between teaching and testing. Despite the novelties in the understanding of teaching and learning a language, testing is still structural and traditional. Thus, Morrow (1979, cited in Heaton, 1990) called for an urgent need in order to use the language for real purposes in context.

As a result, communicative approach came to the ground in which how language is used in communication is of paramount importance. It shares some similarities with integrative approach because both give more importance to meaning than form. On the other hand, they are totally different. In communicative approach, real life tasks are utilized to test the language skills of learners, and the aim is to measure how learners use language while communicating in real-life situations or tasks. Apart from a mastery of the grammar of a language, communicative competence is also needed for successful and effective communication. Communicative tests should reflect the culture, as the focus is on context and the use of authentic materials. Before that approach, for Brown (2003), the problem with language testing was that the tasks were artificial and not reflecting the real use of the language. With the introduction of communicative approach, authenticity is taken as the core to be tested in language. The term assessment has started to be used by the scholars with this approach (Brown, 2003). In a similar vein, Heaton (1990) stated that with the introduction of communicative approach, the concept of qualitative assessment rather than purely quantitative has come out. With the advent of communicative approach, "authenticity of tasks and genuineness of texts" have gained importance (Brown, 2003, p. 11), and "performance assessment found a rationale in the theory of communicative competence" (McNamara, 1997, p. 131).

Furthermore, there have been current issues recently regarding language assessment that are alternative assessment and computer-based testing (Brown, 2003). It is seen that traditional assessment, including pen and paper tests, which mainly relies on scores cannot be a sole indicator of student progress and achievement. Mertler (2003) expressed that traditional assessment is easy in terms of scoring because it has only one true answer which makes scoring easy and fair for teachers. However, with the changes in education, there has been a shift towards more "hands-on, experiental learning" (p. 5). This has led to the advent of alternative assessment. Due to this, alternative assessments have become popular, which have come out as a supplement to traditional assessment because alternative assessment is more authentic and requires a meaningful context (Brown, 2003). Mertler (2003) in his book divided alternative assessment into three subgroups that are informal assessment, performance-based assessment, and portfolio assessment. Informal assessments include questions and observations of teachers which are very often done by teachers during teaching and learning process (Mertler, 2003). Performance-based assessment is an integration of "oral production, written production, open-ended

responses, integrated performance, group performance, and other interactive tasks" (Brown, 2003, p. 11) rather than giving paper and pen responses. Interactive task is at the heart of performance-based assessment. Learners are expected to perform the behavior which is going to be assessed by the teacher. In performance assessment, learners are engaged with real-world tasks; hence, it is more learner-centred. With the help of interactive tasks, learners' performances are assessed. Portfolio assessment is the last type that is defined as simply a collection of student writing over a time, indicating the stages learners have undergone in this process (Hamp-Lyons, 2006). As learners and teachers work on portfolio together and collaboratively, teachers can have a better idea for the assessment of learners, and the level attained by them. Thus, "portfolios are a tool for thoughtful classroom assessment" (p. 154), and they have the potential to increase student learning, and provide a self-assessment for learners (Hamp-Lyons, 1996).

With the use of alternative assessments, assessment for learning has gained utmost importance in educational contexts (Black & Wiliam, 1998). This understanding has been recently mentioned in dynamic assessment (Lantolf, 2009) which is rooted in Vygotsky's term "zone of proximal development" referring to the gap between what a learner can do now and the target production (Fulcher, 2012). Vygotsky pointed out that rather than the outcome of development, process should be investigated and given priority (Lantolf & Thorne, 2006). Thus, giving importance to process rather than product makes dynamic assessment different from traditional assessment practices. According to dynamic assessment, instruction and assessment should go hand in hand, and through dynamic assessment, teachers can have the chance to assess learners during instruction by investigating what they can do alone and with assistance.

Another current trend in testing is computer-based testing in which learners give their responses on a computer. Computer-assisted or web-based tests are the other names for this type of test (Brown, 2003). The most popular type of computer-based testing is computer-adaptive tests in which a learner is given some questions by the computer that are often in accordance with the level of the learner (Brown, 2003). After the computer scores the learner's responses, it is the computer that decides on the next question to be asked based on the previous answers of the learner. As the learner gives correct responses, the computer asks more difficult questions for the next time. Gruba and Corbel (1997) stated that computer-adaptive tests have many advantages such as "reduced administration time, decreased candidate frustration, self-paced tests, the production of

immediate results, a need to have fewer test administrators, and improvements in test security" (p. 141). This type of testing is advantageous for the individualization and self-directed testing; however, there exist some disadvantages as well such as the scarcity of open-ended items, ease of cheating, lack of security, and the abundancy of multiple choice items (Brown, 2003).

The history of testing and assessment mentioned above indicate the stages of testing and assessment, the popularity of testing in the history and the advent of assessment as a complement to testing.

1.3. Importance of Language Assessment

Good assessment practices are crucial because the quality of the assessments that are utilized is a prerequisite for the quality of the instruction and learning (Stiggins, 1999). Purpura (2016) pointed out that the reason why language assessment is necessary is that it helps elicit learners' L2 performance, and with the help of this, certain scores, descriptions or notes are gathered based on the performance of learners, and they are used to make decisions about learners. Shepard (2000) stated that through good assessment practices, more valid decisions can be made in order to adapt instruction and appeal to learners' needs more. Thus, as Marzano (2000, p. 21) indicated for the betterment of student achievement, comprehension and the application of effective assessment practices are crucial. Brown (2003, p. 16) touched upon the importance of language assessment in seven items:

1. Periodic assessments, both formal and informal, can increase motivation by serving as milestones of student progress.
2. Appropriate assessments aid in the reinforcement and retention of information.
3. Assessments can confirm areas of strength and pinpoint areas needing further work.
4. Assessments can provide a sense of periodic closure to modules within a curriculum.
5. Assessments can promote student autonomy by encouraging students' self-evaluation of their progress.
6. Assessments can spur learners to set goals for themselves.
7. Assessments can aid in evaluating teaching effectiveness.

Thomas, Allman, and Beech (2004) drew the attention to the importance of good assessment practices by stating that not only teachers but also students make use of good assessment practices. These assessment practices are good informants; so, with the help

of these, teachers can adapt the pace of the lesson, make a decision about whether the course content is relevant or not, shape student learning during teaching process, get an idea about whether the teaching is effective or not, and help create a confidence in students for the national standardized tests.

1.4. The Role of Teachers in Language Assessment

Assessment covers a wide range of assessment activities such as developing paper-pencil tests, grading, and interpreting the results (Zhang & Burry-Stock, 2003). A language teacher has this assessment responsibility as a part of her/his profession (Mertler, 2003). As teaching and assessment are the concepts affecting each other, they inform and improve each other (Malone, 2013); thus, teachers have great roles in bridging between these two concepts. The role of teachers is made salient in assessment process with the utterances of many scholars in the literature (Stiggins, 1999; Popham, 2009) who pointed out that when teachers have the necessary knowledge and skills for assessment, it becomes more possible to talk about effective assessment activities. With this great role in language assessment, teachers' knowledge of assessment has a big impact on the quality of education (Malone, 2013). Regarding this, Calderhead (1996) stressed that the power of assessment relies on the knowledge and practices of teachers. As a result, it is necessary for teachers to utilize assessment strategies to make decisions, to decide on the most suitable instruction for learners, and to get an idea about teaching and learning progress. In other words, effective teachers are conscious about what, how, and why they are making use of assessment practices (Stanford & Reeves, 2005).

Though the literature indicates the centrality of teachers in language assessment and the important roles teachers play in this process, Mertler and Campbell (2005) drew attention to a problem which reveals that most of the teachers do not think that they are ready for their roles in assessment. Alderson (2005, p. 4) touched upon this issue by stating that "tests made by teachers are often of poor quality, and the insight they could offer into achievement, progress, strengths and weaknesses is usually very limited indeed". These utterances above indicate the crucial roles of teachers in assessment, and the vital need for assessment literacy which is a must to be able to assess learners and teaching-learning process effectively.

1.5. Assessment Literacy

Taylor (2013) stated that the meaning of literacy which is "being able to read and write" has expanded for the last decades, and the notion of literacy has been used with various concepts such as academic literacy, media literacy, etc. No matter which concept it is used with, the meaning does not change, and the "focus is on the ability to understand the content and discourse associated with a given domain or activity and on being able to engage with and express oneself in relation to this" (p. 405). One of the concepts used with literacy is the assessment, and the term "assessment literacy" was coined by Stiggins in 1991. The American Federation of Teachers, the National Council on Measurement in Education, and the National Education Association (1990, pp. 31-32) came up with seven standards that are crucial for Teacher Competence in the Educational Assessment of Students. These standards include:

1. choosing assessment methods appropriate for instructional decisions,
2. developing assessment methods appropriate for instructional decisions,
3. administering, scoring, and interpreting the result of both externally-produced and teacher-produced assessment methods,
4. using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement,
5. developing valid pupil grading procedures which use pupil assessments,
6. communicating assessment results to students, parents, other lay audiences, and other educators,
7. recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information.

These standards are the features an assessment literate teacher is expected to possess. Though a lot of people support assessment literacy, there is not an agreement on the exact definition of this term (Fulcher, 2012); thus, scholars define their own understanding of assessment literacy in various terms. DeLuca, Valiquette, Coombs, LaPointe-McEwan, & Luhanga (2016, p. 1) drew the attention to the importance of this term by stating that assessment literacy is the core of the professional development. Stiggins (1995, p. 240) defined it as "knowing the difference between sound and unsound assessment". For Falsgarf (2005, p. 6), assessment literacy "is the ability to understand, analyze, and apply information on student performance to improve instruction". Mertler and Campbell (2005, p. 16) defined it as "teachers' knowledge and abilities to apply assessment concepts and techniques to inform decision making and guiding practice". Another definition belongs to Mertler (2003) for whom, assessment literacy is the key

element between the quality of assessment and learner achievement. Besides, Stiggins (1995, p. 240) emphasized that assessment literate teachers know "what they are assessing, why they are doing it, how best to assess the skill, knowledge of interest, how to generate good examples of student performance, what can potentially go wrong with the assessment, and how to prevent that from happening".

As is clear, assessment literacy covers the knowledge related to assessment and also application of this knowledge during assessment practices. Assessment literacy was first introduced in general education, and then became familiar in language education. That's why, most of the research studies have been in the field of general education and psychology. Very recently, a new term, language assessment literacy, has flourished, and a new research area has come out in language education as well.

1.6. Language Assessment Literacy

Language assessment literacy is rooted in the term assessment literacy, but it has appeared as a distinct area from assessment literacy. Malone (2013, p. 329) defined language assessment literacy as "language teachers' familiarity with testing definitions and the application of this knowledge to classroom practices in general and specifically to issues related to assessing language". For Inbar-Lourie (2008, pp. 389-390), "language assessment knowledge base comprises layers of assessment literacy skills combined with language specific competencies, forming a distinct entity that can be referred to as language assessment literacy". As is seen, language assessment literacy requires additional competencies related to language when compared to assessment literacy. However, this field and language assessment literacy concept are very novel, and research into language assessment literacy "is still in its infancy" (Fulcher, 2012, p. 117). The studies related to language assessment literacy is very rare, and they mostly focus on the needs of language teachers in relation to language assessment (Inbar-Lourie, 2008; Fulcher, 2012; Malone, 2013).

1.7. The Importance of Teachers' Language Assessment Literacy and Language Assessment Knowledge

Teachers are the key elements in the process of assessment (Leung, 2014). As they are crucial in this process, Popham (2006) stated that it is vital for teachers to have a certain degree of assessment literacy, because teachers are engaged in assessment-related

activities during teaching and learning process. For language teachers, it is a must to have language assessment literacy. The reason behind this is that it is the teacher who is responsible for developing assessment methods, administering, scoring and interpreting assessment results, developing grading procedures, communicating assessment results, and using them in making educational decisions (Stiggins, 1999). Furthermore, all these processes related to assessment require a teacher who has adequate knowledge and abilities in assessment (Alkharusi, 2011). To put it simply, a teacher who does not have sufficient assessment knowledge cannot be good at developing tests, administering, scoring and interpreting them.

It is crucial that teachers have a certain degree of language assessment knowledge as a part of language assessment literacy. In other words, language assessment knowledge is the core of language assessment literacy which is the combination of knowledge and skills related to language assessment. Language assessment knowledge is a must, because all activities of teachers associated with assessment such as developing a test, scoring, administering, and interpreting it in separate language-related areas, depend on the teachers' language assessment knowledge. Without having adequate language assessment knowledge regarding language skills that are reading, writing, speaking, and listening, it is not very possible for a teacher to be effective in assessing language-related skills.

1.8. Statement of the Problem

It is crucial that language teachers need to have adequate knowledge in assessment-related process (Price, Rust, O'Donovan, Handley, & Bryant, 2012). However, many inservice teachers stressed that they are not adequately equipped with assessment knowledge (Plake, 1993). Stiggins (2010, p. 233) pointed out this problem with a very assertive utterance by stating that "assessment illiteracy abounds". This indicates that teachers are responsible for assessing learners, but whether they have the necessary knowledge to assess learners is open to discussion. To help learners achieve higher levels, teachers should have high level of assessment literacy (Coombe, Davidson, O'Sullivan, & Stoynoff, 2012). Popham (2009) expressed that teachers are expected to assess learners' proficiency and progress, but many teachers do not have sufficient understanding associated with very basic terms in assessment. As the literature on assessment literacy shows that teachers are not competent enough in assessment

knowledge, how teachers can assess learners' proficiency and progress efficiently is under discussion.

In addition to these, as Purpura (2016, p. 191) stated "rather than seeing assessment as an organic part of applied linguistics, L2 assessment is still often viewed as an afterthought, or as a craft". As assessment does not get its deserved interest, it is better to approach this issue from the perspective of language teachers who are responsible for the assessment of learners. Language teachers as assessors should be explored first in terms of their assessment knowledge, because the whole assessment process is based on the competency of language teachers in language assessment. If language teachers are not competent enough in language assessment, the whole assessment process will be misled by the incompetency of these language teachers.

In Turkey, it is language teachers' duty to assess learners with the help of various kinds of assessment practices through standardized and classroom-based tests. Through these various ways of assessment practices, learners' proficiency in English is assessed by language teachers. Language teachers are responsible for assessing learners through formal and informal assessment practices. Thus, they are expected to develop tests, administer and score them, and interpret the results of assessment practices. The problem is that the expectation from them is big, but they do not have much exposure to training in assessment. In pre-service education, there is only one course related to testing and assessment, and pre-service teachers are not expected to assess learners during practicum. What is more, it is not obligatory for language teachers to participate in professional development programmes regarding assessment. Hence, the result is that language teachers have only one course related to assessment, and this course covers basic terms related to assessment in general, and assessing language skills in particular, and training is not a must.

Indeed, Higher Education Council (HEC) determined two courses in undergraduate programmes of ELT. One is assessment and evaluation in education, and the other one is English language testing and evaluation. However, there is no specification or framework for the contents of these courses. Because of this, the undergraduate programmes of three leading universities in Turkey were analysed in detail to learn the contents of these courses. It was seen that even the name of the course was different for each university. Moreover, one university was in line with the programme of HEC and there are two courses related to assessment, however, in the others, there is only one course in

undergraduate programme. Additionally, the contents of these courses are highly different from one another. For instance, one university focuses on the importance of assessment, basic terminology, subjective-objective assessment, features of a well-developed exam, test usefulness, reliability, validity, types of exams, TOEFL, assessing general language proficiency, features of a standard proficiency exam, assessing skill development, assessing reading, speaking, writing and listening, and designing tests. The other university focuses on the types of tests, test preparation techniques for the purpose of measuring various English language skills, the practice of preparing various types of questions, evaluation and analysis techniques and statistical calculation. The last university focuses on theories of measurement, evaluation and assessment construction and evaluation of tests for assessing second language skills, reliability, validity, backwash, statistical analysis of test score, and interpretation of results. The contents may seem comprehensive at first sight, but all these topics have to be covered within just one academic term which is too short to learn how to assess language skills and make practice based on the theories. Besides, assessing language skills and statistical calculation should be separated from each other, because individually they are already very detailed, and it is even impossible to cover all the topics mentioned in the course specifications in just one term. The problem gets worse when language teachers who are not graduates of English language teaching department assess learners, because these language teachers do not even have this course in pre-service education.

With this insufficient background in assessing language skills, graduates start to work as English language teachers. Preparatory programmes of the universities are one of the workplaces where teachers are expected to teach English, and assess their learners in each skill. These programmes were purposefully selected for this study because they are the contexts in which each language skill is given importance, and as a result, all skills are assessed. The assessors in these preparatory programmes are language teachers. The problem is that language teachers are responsible for all the assessment-related activities in preparatory programmes, but as is clear, their background in language assessment is not very good. Thus, how knowledgeable or competent they are in assessing learners is the question. As a starting point, language assessment knowledge of language teachers should be determined. However, in Turkey, there is not a study on identifying the language assessment knowledge of language teachers; thus, there is paucity of research in language assessment literacy to shed light on this issue. This identification is vital

because by detecting the strengths and weaknesses of language teachers, the needs of language teachers could be specified. Based on these needs, testing and assessment course in pre-service education and teacher professional development programmes related to language assessment can be designed and developed.

Thus, there is an urgent need in Turkey to explore language assessment knowledge of language teachers in preparatory programmes in order to detect their strengths and weaknesses in language assessment.

1.9. The Purpose of the Study

Referring to the aforementioned need for research, this study aims to focus on the language assessment knowledge of language teachers working at Turkish higher education setting. Due to the paucity of research in this area of inquiry and lack of research instrument to measure this knowledge, one of the primary aims of this study is to develop and validate a scale exploring language assessment knowledge of the language teachers based on a systematic scale-development process.

After developing the scale, the second aim is to get an insight about the language assessment knowledge level of these teachers and investigate their language assessment knowledge with regard to the components of a language that are reading, listening, writing and speaking. Their language assessment knowledge is investigated based on certain demographic features such as years of experience, educational background, the BA programme they graduated from, working at a state or private university, having a separate testing course in pre-service education, attending a professional development programme on testing and assessment and being a member of testing office. Additionally, whether their LAK level changes according to their perceived self-competency in assessing each skill is investigated. As a result, whether language assessment knowledge of language teachers differs based on these demographic features are revealed.

Another purpose of the current study is to demonstrate the participants' opinions regarding their LAK level and the findings of the scale, and the last purpose is to reveal their needs in language testing and assessment.

1.10. Research Questions

As the present study investigates language assessment knowledge of EFL teachers in higher education in Turkey with the help of a scale that was developed and validated

by the researcher, it is aimed to find out answers to the following research questions throughout the study:

1. What are the psychometric properties of Language Assessment Knowledge Scale (LAKS)?
2. What are the general and skill-based Language Assessment Knowledge (LAK) level of EFL teachers in Turkish higher education setting?
3. Is there a relationship among their levels of skill-based LAK?
4. Does LAK level change according to following demographic features
    - years of experience,
    - educational background,
    - the BA programme being graduated,
    - workplace,
    - having a testing course in BA,
    - attending trainings on testing and assessment and
    - being a testing office member?
5. Does their LAK level change according to their perceived self-competency in assessing each language skill?
6. What are the opinions of EFL teachers in Turkish higher education setting regarding their LAK level and the findings of the scale?
7. What are their needs in language testing and assessment?

1.11. Definition of Key Terms

Testing: a method of measuring a person's ability, knowledge, or performance in a given domain (Brown, 2003, p. 3)

Assessment: a procedure for eliciting test and nontest for the purpose of making inferences (Purpura, 2016, p. 191)

Assessment includes testing, and all tests are formal assessments (Brown, 2003).

Language assessment: a broad term referring to a systematic procedure for eliciting test and nontest for the purpose of making inferences or claims about certain language-related characteristics of an individual" (Purpura, 2016, p. 191).

Language assessment literacy: language teachers' familiarity with testing definitions and the application of this knowledge to classroom practices in general and specifically to issues related to assessing language (Malone, 2013, p. 329)

Language assessment knowledge: having knowledge about and being familiar with basic terms, concepts and ways of assessing language skills that are reading, listening, writing and speaking, and also having knowledge related to designing tests, administering and scoring them based on these four skills.

2. LITERATURE REVIEW

2.1. Assessment and Teaching

Teaching and assessment cannot be thought separately, because assessment is a component of learning and teaching process, and teachers are engaged in assessment and assessment-related activities in most of their professional time. They cannot be separated, because "assessment and instruction are two sides of the same coin" (DiRanna, et al. 2008, p. 22). Instead of seeing assessment something external to the instruction, it is better for teachers to regard assessment as an integral part of instruction, and such a point of view provides teachers "a window onto classroom learning processes so as to be able to measure and track students' language growth, encourage learner engagement in the learning process, and determine the appropriacy of instruction in meeting students' learning needs" (Katz, 2012, p. 66).

However, assessment can sometimes be confusing for many people because of the presence of various terms in relation to assessment, which are assessment, test, measurement and evaluation. These terms abound in the literature, and they all refer to different things. As Brown (2003) mentioned, assessment is the general term covering tests, and tests are the systematic procedures. Tests are prepared tasks; on the other hand, assessment is more than tests. Additionally, assessment can cover any instructional activities including tests; in other words, test is a subset of assessment. The difference between them was made clear in the following sentence: "assessment refers to a broad array of methods and approaches to collect information so as to make decisions about learning in contrast to the term testing, which is used to refer to one form of assessment" (Katz, 2012, p. 66). Evaluation, the next term, is related to making judgments related to a person or a thing; thus, worth is included in evaluation (Brady & Kennedy, 2014). The final one, measurement is "the systematic classification of observations of student performance" (Brady & Kennedy, 2014, p. 171). The confusion between evaluation and measurement was discussed by McInerney (2014) as follows: measurement refers to the scores a learner gets from a test; on the other hand, evaluation "refers to the quality, value, or worth of the information gathered (Gronlund, 1985; Mager, 1990a, 1990b)" (p. 315). To give an example for the difference between them, imagine that there is a child who weighs 15 kilograms, which is measurement (an objective measurement). If that child is thought to be skinny, then it is evaluation (a subjective/evaluative judgment) (McInerney, 2014).

Assessment is a crucial issue for policymakers, education and public because it can have an effect on educational reform due to its power to guide educational development (Cromey & Hanson, 2000). To draw attention to the importance of assessment, Plake (1993) stated that teachers spend 50% of their professional time on assessment-related activities. Mertler (2009) also expressed that assessing learners is one of the most important and critical duties of teachers. It is crucial because when teachers can understand and make use of effective assessment practices, they get closer to increasing learner achievement (Marzano, 2000).

Coombe, Troudi, & Al-Hamly (2012, p. 20) discussed assessment through the lenses of various partners of assessment who are all included in the process of assessment. In terms of students, assessment is both an instructional activity and also the prediction about teachers' expectations (McLaughlin & Simpson, 2004). Many students perceive assessment not in relation to themselves, but in relation to their teachers. For them, assessment is something teachers do, and they are not actively engaged with assessment but they are just somehow influenced by teachers' assessment-related activities. In the eyes of the students, assessment is also a factor leading them to stress and nervousness, and they experience test anxiety; thus, because of assessment they feel under pressure. For teachers, assessment is not perceived very differently from students. Teachers relate assessment to unpleasant feelings. Teachers think that there exists a gap between assessment and instruction, and most tests are developed by teachers who have no classroom experience, which is a big problem for many teachers. Jacobs and Chase (1992) came up with a conclusion after their research study that assessment-related activites are one of the unpleasant duties of teachers, and many teachers are not very content with the implementation of assessment-related activities. The last point of view belongs to educational boards. Each and every standard for teachers includes assessment as one of the main points.

Many assessment competencies have been put forth by the National Education Association (NEA), the American Federation of Teachers (AFT), the Council of Chief State Schools Officers (CCSSO), the National Council for Accreditation of Teacher Education (NCATE), and the National Board of Professional Teacher Standards (NBPTS). In a broader field, TESOL and the National Council for the accreditation of Teacher Education (NCATE) developed the TESOL/NCATE standards for teacher education. The standards are composed of five main domains one of which is assessment.

Hence, it is obvious that the necessity and importance of assessment are regarded as crucial around the world.

As is clear from aforementioned discussion, assessment is crucial in instruction, and an indispensable part of it. As assessment has an important place in instruction, there exist many reasons why assessment is carried out. Based on why assessment is carried out, the followings are the types of assessment in the literature. Depending on the purpose of assessment, the kinds of assessment abound in the literature, which are as follows:

2.1.1. Formative assessment and summative assessment

Brown (2003) stated that assessment can take two forms as formative and summative assessment regarding its function, that is, how the assessment is going to be used. In formal assessment, the primary focus is on the continuous development of the learner. All informal assessments are formative assessments, because the aim here is to foster the learning of learners. However, in summative assessment, the purpose is "to measure, or summarize, what a student has grasped, and typically occurs at the end of a course or unit of instruction" (p. 6). For Popham (2009), summative assessment helps teachers decide "go/no-go decisions based on the success of a final-version instructional program" (p. 5); on the other hand, formative assessment is used by teachers to adjust the programme and by students to adjust their learning. Improvement is the core and center of formative assessment. Quizzes can be considered as a form of formative assessment when they are carried out to see how effective the instructional programme is for the teacher, and how successful the students are. For Katz (2012), formative assessment sees learners as collaborators in educational process, and this kind of assessment aims "to monitor and support student learning and to fine-tune instruction so that it meets students' evolving needs" (p. 67).

2.1.2. Informal assessment and formal assessment

For Brown (2003), informal assessment can take place in different formats such as asking a question to a learner or just putting a smile on an assignment. It can sometimes be a comment; thus, informal assessment can be unplanned and spontaneous. On the other hand, formal assessment is designed specifically, and it is usually systematic and planned. Popham (2009) considered formal and informal assessment as a subgroup of formative assessment by naming both of them as classroom assessment. For Popham, formative

assessment is utilized when teachers want to know what their students know and their potential of what they can learn more, when teachers need a change in the programme, and when they want to see whether the activities they are using are appropriate and effective or not for the learners.

2.1.3. Direct assessment and indirect assessment

As Hughes (1989) mentioned, based on test construction, assessment is divided into two as direct and indirect assessment. Direct assessment requires the learner to be engaged in the activity directly; thus, there is the actual performance of the learner; however, in indirect assessment, learners are not actually performing the target task (Brown, 2003). In indirect assessment, the abilities which underlie the skill we want to test are measured (Hughes, 1989). For Hughes (1989), direct assessment has some advantages. To start with, it is straightforward and to the point as long as the abilities that are assessed are clear. Secondly, assessing and interpreting learners' performance are straightforward in productive skills. The last one is that it has a beneficial washback effect provided that the test includes the skills which the tester wishes to foster. Indirect assessment has also certain plusses (Hughes, 1989). It offers more possibilities of including more samples than direct assessment, which is crucial for increasing reliability. For example, asking learners to write to assess their writing skills is an example of direct assessment, but if you wish to assess writing skills by indirect assessment, then you have more chance to include more samples, which in turn leads to greater reliability. Despite this advantage, indirect assessment is questioned because of blurred relationship between performance on them and performance of the skills; in other words, it is not very easy and clear to assert that the underlying skills that are measured are a way of predicting the writing ability of learners (Hughes, 1989). Additionally, Davies (1999) stated that since the relationship between the test performance and future use is stronger in direct assessment, face validity is higher in direct assessment when compared to indirect assessment.

2.1.4. Objective assessment and subjective assessment

As Hughes (1989) stated, the difference between objective and subjective assessment stems from the scoring procedure. In an objective assessment, "correctness of the test taker's response is determined entirely by predetermined criteria so that no

judgment is required on the part of the scorer"; however, in a subjective test, the scorer's interpretetaion determines the correctness of the response (Bachman, 1990, p. 79). A multiple choice test is considered as the most common form of objective assessment, because the rater's judgment is not included in the scoring process. The degree of subjectivity may change. For instance, the subjectivity in scoring a composition is different from the scoring a short answer item. The degree of subjectivity in scoring a composition is greater than the short answer item, and ensuring objectivity is a prerequisite for a reliable test.

2.1.5. Discrete point assessment and integrative assessment

Discrete point assessment is the assessment of one element at a time, item by item (Salim, 2001, p. 179). On the other hand, in integrative assessment, there is the combination of many elements. For instance, assessment of a particular grammatical structure is discrete point assessment; on the other hand, asking learners to write a composition, taking notes while they are listening, taking a dictation or completing a cloze passage are integrative ways of assessment (Salim, 2001, p. 179). For Aslam (1992), integrative assessment is rooted in cognitive view of language in which language is regarded as a whole, and language is more than its parts. Owing to this, skills, aspects or levels are not separated in integrative assessment as in discrete-point assessment.

2.1.6. Norm-referenced assessment and criterion-referenced assessment

Brown (2003) explained the difference between norm-referenced and criterion-referenced as follows. Norm-referenced assessment focuses on the mean score, standard deviation, median and percentile rank of the learner; thus, it requires mathematical calculations. To be able to interpret a learner's score, aforementioned terms (e.g. mean) are needed. There is a rank order, and the learner is placed along the continuum based on the score. However, criterion-referenced tests "are designed to give feedback, usually in the form of grades, on specific course or lesson objectives" (p. 7). For Shrock and Coscarelli (2007), norm-referenced assessment includes items that divide the test takers' scores from each other. However, in criterion-referenced assessment, the items rely on specific objectives. To put it differently, the performance of a learner depends on the other learners in norm-referenced assessment; on the other hand, this performance depends on the specific criteria in criterion-referenced assessment. Thus, in criterion-referenced

assessment, there is no limit in the number of the learners who can be regarded as successful; but, in norm-referenced assessment, the number of the learners who can be regarded as successful is limited.

To sum up, based on the objectives of assessment, assessment can take many forms such as formative vs summative, informal vs formal, direct vs indirect, objective vs subjective, discrete-point vs integrative, and norm-referenced vs criterion-referenced assessment.

2.2. Principles of Language Assessment

Each and every assessment should meet certain criteria to be considered as a good test. There are some properties of a good test that are validity, reliability, practicality, washback and authenticity (Harris, 1969; Hughes, 1989; Brown, 2003; Farhady, 2012). It is not very easy and simple to develop a test, but it is a necessity to take into consideration these principles if the purpose is to develop an acceptable and defendable test (Farhady, 2012, p. 45).

Below are the principles of a good test:

2.2.1 Validity

Validity is seen as the central concept in testing and assessment (Fulcher & Davidson, 2007), because a valid test means "a good test" (Fulcher & Davidson, 2012, p. 21). For Oller (1979), validity is the most important quality a test has to have, because validity is the basis of a test. There are many definitions of validity in the literature. The followings are some of them: Henning (1987, p. 170) defined it as "appropriateness of a given test or any of its component parts as a measure of what it is purposed to measure". For Hughes (1989, p. 22), "a test is said to be valid if it measures accurately what it is intended to measure". Harris (1969) also stated that there are two questions to be asked to ensure validity which are (1) what precisely does the test measure? and (2) how well does the test measure? (p. 19). For Gronlund (1998, p. 226), validity is "the extent to which inferences made form assessment results are appropriate, meaningful, and useful in terms of the purpose of the assessment". As an example given by Akbari (2012), think of a test that is designed for measuring vocabulary through analogies. If that test measures general intelligence instead of vocabulary, then this test has no validity, because it does not measure the intended ability or skill. The followings are the kinds of validity.

2.2.1.1. Content validity

If the content of a test includes a representative sample of skills, structures, etc., then it has content validity (Hughes, 1989). Akbari (2012) uttered that content validity "checks the representativeness of a test content to make sure content sampling has been carried out in a theoretically justifiable manner" (p. 31), and added that the selection of the content is not a difficult task in achievement tests because the test has a domain of the curriculum or the coursebook that is determined in advance. For Akbari (2012), the risk of inadequate content always exists, since the content may be appropriate but may be inadequate in terms of coverage and relevance. Asking for expert opinions is a way of having content validity because expert opinions lead testers to decide on the representativeness of the samples in the test (Fulcher & Davidson, 2007). For content validity, what should be done is the careful and detailed investigation of the related skills and structures, and then the items in the test should be representative of all the skills and structures covered, and the items should "represent adequately each portion of the analysis and outline" (Harris, 1969, p. 19).

Brown (2003, p. 23) gave two instances for content validity. First one is that if a course has ten objectives, but the test measures only two of the objectives, here there is a threat to the content validity of the test. Second one is that a test aims to measure the speaking skills of the learners. When the learners are required to speak in a given context, then it is okay in terms of content validity. On the other hand, if the learners are given a paper and pencil multiple choice test for assessing their speaking skills, then content validity of the test suffers. When certain areas are not tested, they will not be given enough importance during teaching and learning process; hence, the best solution is to come up with test specifications first, and then make it sure that the test should include and reflect these (Hughes, 1989).

2.2.1.2. Construct validity

For Bachman and Palmer (1996, p. 21), construct validity is related to the "meaningfulness and appropriateness of the interpretations that we make on the basis of test scores", and the main question to be asked for a test to have construct validity is "to what extent can we justify these interpretations?". The exact definition of what is meant by construct is an important issue (Fulcher & Davidson, 2007). Brown (2003, p. 25)

expressed that constructs do not have to be directly or empirically measured, and "their verification often requires inferential data". To give an example, proficiency is a linguistic construct, and self-esteem is a psychological construct. For Akbari (2012), construct validity is not only abstract but also empirical. It is regarded as abstract because there should be a theory of proficiency or skill it aims to measure, and it is regarded empirical since "it must be checked statistically against that theory through highly sophisticated statistical techniques" (p. 32). If a test measures each and every skill individually or measures them as a unit, it is directly related to construct validity.

2.2.1.3. Criterion-oriented validity

The logic behind criterion-oriented validity is that "if a test accurately measures a certain component or skill of the L2, it should closely correlate with other tests that measure the same component or skill" (Akbari, 2012, p. 31). The focus of the tester is on the link between the test and the criterion on which the tester is going to make predictions (Cronbach & Meehl, 1955, cited in Fulcher & Davidson, 2007). Harris (1969) mentioned this type of validity with a different name, "empirical validity", but intending the same idea. For Hughes (1989, p. 23), criterion-validity is defined as "how far results on the test agree with those provided by some independent and highly dependable assessment of the candidate's ability", and this "independent assessment is the criterion measure against which the test is validated". Harris (1969) said that when the correlation between test scores and the external or independent criterion that is seen as trustworthy is high, then this test is said to have criterion-oriented validity. As an example for criterion-oriented validity given by Akbari (2012), a teacher developed a vocabulary test, and that teacher correlated the results of the vocabulary test with the results of a well-known high-stakes test. These two tests were both given to the same group of learners, and the purpose of the teacher here is to see the degree of correspondence of both tests. The logic behind this is the comparison of both tests.

There are two types of criterion-oriented validity, which are concurrent validity and predictive validity. Brown (2003) differentiated these types as in the following statements: When the other concurrent results or performances are in line with the score or the performance of the current test, then it can be said that this current test has concurrent validity. Predictive validity has nothing to do with the other concurrent tests. If a test can predict the probability of the test taker's future success, then this test has

predictive validity. Akbari (2012, p.32) expressed that "in concurrent validity, both the newly developed test and the criterion test are administered at the same time to a group of test takers and the scores obtained on both are correlated". When the new test has a strong and positive correlation with the criterion test, then there is concurrent validity here. However, in predictive validity, Akbari (2012, p. 32) stated that "there is a time gap between the administration of the newly developed test and the criterion test". If there is strong and positive correlation, then it means that predictive validity exists.

Hughes (1989, p. 25) gave the following examples for concurrent and predictive validity. For concurrent validity, the aim of the test is to measure oral skills of learners as a part of a proficiency exam, but there is a time limit. The objectives are listed and the functions to be included in the oral exam are determined. In order to test all the functions, a 45-minute exam is necessary for each learner, but for such a test that is the component of a proficiency exam, it is not practical. Then, a question comes out: Can a 10-minute oral test measure learners' oral skills? This 45-minute oral test is the criterion against which the 10-minute test will be compared. Randomly selected learners are administered to these both oral tests, and the scores obtained from these two tests are compared to each other. The scorers of the short test are not aware of the results of the longer test. If there is a high agreement between these two tests, then it can be said that the shorter version of the text is valid, because its results are similar to the results of the longer version of the test. An example for predictive validity is to see how well a proficiency test can predict a learner's ability to deal with a course at a different university. The criterion measure here could be the supervisor's comments or the outcome of the course.

2.2.1.4. Face validity

For Farhady (2012), it is "the extent to which the physical appearance of the test corresponds to what it is supposed to measure" (p. 38). In other words, it is how the test-takers, educators, etc. perceive the test, and how it looks (Harris, 1969, p. 21). Harris (1969) added that face validity is important, because if the test-takers do not find the content and the items in the test appropriate, then they cannot adopt the test, which, in turn, affects their motivation negatively. Two examples are given by Farhady (2012) to make it clearer. One is that if learners expect to see multiple choice questions in the test, and the questions in the test are open-ended questions, then this situation will influence the test performance of the learners because of the low face validity of the test. The other

one is that using cloze test to measure grammar may lead to low face validity of the test, because cloze test does not give the message to the learners that their grammar is being measured. Indeed, it is well accepted among the scholars that cloze test is a valid way of measuring grammar competence of learners; however, as the appearance of cloze test does not give that message, the test has a low face validity. Obviously, face validity is important, but Farhady (2012) warned that face validity is an issue which should not be given too much importance. For sure, it should not be ignored at all. What is stated here is that though a test has no face validity, that test can still be regarded as valid.

2.2.2. Reliability

Bachman and Palmer (1996, p. 19) defined it as the "consisteny of measurement", and added that in different test situations, the scores should be consistent. In other words, if the same test is administered to a group of learners in different test situations, there should not be differences between the scores of learners across different test locations and occasions. For Brown (2003), a reliable test is also dependable along with being consistent. Farhady (2012) drew attention to the importance of reliability by giving the following example. Think that there is a learner who got a score of 40 out of 100 items in a grammar exam. That student wanted to increase the score, and took the same test again. For this time, she got 90, and two days later, she got 70. There is something miserable in these results, because it is clear that a student's knowledge of English cannot be changed in such a short period of time (two days). Then, it is certain and clear that there is a problem with the reliability of the test. Farhady (2012) stated that reliability is represented by the letter "r", and it is between "0" as a minimum degree and "1" as a maximum degree. When reliability is of concern, the scores obtained from the test matter, not the form or the content of the test, and the test itself has no meaning and importance to calculate reliability, because what is needed is only scores. Brown (2003, p. 21) explained the following possibilities affecting the reliability of a test.

2.2.2.1. Student-related reliability

Illnesses, tiredness of a learner, all the physical and psychological factors are in this group. Along with these, learners' making use of certain strategies can also be included in this category.

2.2.2.2. Rater reliability

During the scoring, there are many factors such as subjectivity and bias of the raters. It can be divided into two as intrarater reliability and interrater reliability. Intrarater reliability may be affected negatively due to unclear scoring scales, tiredness or biases. For example, if a teacher is to read 100 essays in a couple of days, then because of tiredness there may be fluctuations between the scores of the learners. As an example for interrater reliability, when two or more than two scorers score the same work in a different way, then it is a threat to interrater reliability, because there occur inconsistencies among raters. Lack of attention, inexperience and biases may be the sources of this inconsistency. Harris (1969) stated that rater reliability is great in multiple choice tests; however, it has some fluctuations in free-response tests.

2.2.2.3. Test administration reliability

The conditions in which the test is administered may have a negative effect on the reliability of a test. Some of these conditions might include issues related to photocopying, light, noise coming from outside, and the comfort of the desks. For instance, in a listening exam, when learners sitting near the window have difficulty in hearing the tape recorder, then it is a threat to reliability.

2.2.2.4. Test reliability

The test itself can sometimes cause problems. If the items in the test are not clear enough or they are too long to be answered carefully, then test reliability is under threat. For Harris (1969), test reliability is also associated with the adequacy of samples. When there are more samples of learners' performances, then the test becomes more reliable. Calculation of the reliability requires certain statistical knowledge and analysis. There are four common ways of estimating reliability, which are the test-retest method, parallel forms method, split-half method, and the KR-21 method (Farhady, 2012).

2.2.2.4.1. The test-retest method

Giving the test twice to the same group of learners is a way of seeing if the test is consistent or not. The scores obtained from the first test, and the second test, retest, are compared, and the correlation of them is calculated which gives information about the

reliability of the test. When there is a time interval between the administration of two tests, then it is called as "stability of scores over time" (p. 40). The test-retest method has some drawbacks. To begin with, it is not very possible to have the same group of learners taking the first test in the retest. Besides, when there is time interval between test and retest, there might be other factors having an impact on the scores such as memorization and practice. Along with these, nobody's knowledge remains the same; thus, the learners' state of knowledge in the retest may be different from the one in the first test.

2.2.2.4.2. Parallel forms method

Using parallel forms of the test means using different versions of the test. This alternate test can be equivalent in length, difficulty, time limits, formats, and all other such aspects (Harris, 1969, p. 15). Harris (1969) warned that practice effect is the drawback of this method. Farhady (2012) expressed it is more advantageous when compared to test-retest method, because there is no need for the administration of the test twice. The disadvantage of this method is that it is not very easy to come up with a parallel form of an existing test. There should be certain logical and statistical criteria that two parallel forms of the tests must meet (p. 40).

2.2.2.4.3. Split-half method

Test-retest method and parallel forms method have certain shortcomings; and due to these shortcomings, split-half method was developed. In this method, the test-taker's test is divided into two halves. The scores gotten from the first half and the second half are calculated and compared. If there is a high correlation between these two scores, then the test is reliable. In this method, two halves are assumed to be equal. Harris (1969) expressed that this division of the test into two is generally done by separating odd- and even-numbered items.

2.2.2.4.4. The KR-21 method

Kuder and Richardson (1937) developed some formulas to be used in testing, and one of the them is KR-21, which is used to "estimate the reliability of a single test given to one group of examinees in a single administration" (p. 41). For this, calculating mean and variation is necessary.

2.2.3. Practicality

When a test is administered, there should be certain features which must be kept in mind under the category of practicality, which are economy, ease of administration and scoring, and ease of interpretation (Harris, 1969). For Farhady (2012), all the facilities that are related to development, administration and scoring procedures of a test are within practicality. For instance, asking learners to write an essay is a valid way of assessing learners' writing skills. However, when a large number of learners are asked to write an essay, then it is not practical. Brown (2003) stated that an effective test is practical, that is, this test is not very expensive, adheres to certain time limits, is easy to administer, and has a time-efficient scoring procedure. Farhady (2012) warned that low and high practicality factors should be taken into consideration while developing and scoring. For instance, developing a multiple choice test is hard, but it is easy to score it. On the other hand, developing an essay is easier compared to multiple choice test; however, it takes more time to score it and is more difficult. Thus, the practicality of the procedures during the development and scoring of different test types may differ.

2.2.4. Washback

There are many definitions of washback in the literature. For Cohen (1994, p. 41), washback is "how assessment instrument affects educational practices and beliefs". Messick (1996, p. 241) defined it as "the extent to which the introduction and use of a test influences language teachers and learners to do things they would not otherwise do that promote or inhibit language learning". Another definiton belongs to Hughes (1989, p. 1) who, with a more specific sense of washback, defined it as "the effect of testing on teaching and learning", and it can be harmful and beneficial. Messick (1996, p. 241) gave an example for beneficial washback as in the following. A language test has beneficial washback when the language tests include "authentic and direct samples of the communicative behaviors of listening, speaking, reading, and writing of the language being learnt". An example for harmful washback could be as follows (Hughes, 1989). Learners practice writing skills through multiple choice items, and they are not expected to write in the test. This situation leads the learners to practice writing skills only through multiple choice items, and these learners do not write anything for the preparation for the test. Here, there is the harmful impact of the test on teaching and learning. Bailey (1996) has a different division and differentiated washback as positive and negative based on the

fact that they foster or impede the realization of educational goals by the learners. Even though the names for groups are different in these classifications, logic is the same. However, Alderson and Wall (1993) have a different view regarding washback, and they said that washback is used when learners and teachers feel obliged to do things they normally would not do due to tests. Based on this definition, washback can help learners give more importance to certain parts because they think that they are going to be responsible for those in the test (Wall, 2012). Alderson and Wall (1993, pp. 120-121) analyzed washback in detail in many research studies around the world, and came up with 15 hypothesis that are as follows:

H.1. A test will influence testing.
H.2. A test will influence learning.
H. 3. A test will influence what teachers teach.
H. 4. A test will influence how teachers teach.
H.5. A test will influence what learners learn.
H. 6. A test will influence how learners learn.
H. 7. A test will influence the rate and sequence of teaching.
H. 8. A test will influence the rate and sequence of learning.
H. 9. A test will influence the degree and depth of teaching.
H. 10. A test will influence the degree and depth of learning.
H. 11. A test will influence attitudes to the content, method, etc., of teaching and learning.
H. 12. Tests that have important consequences will have washback.
H. 13. Tests that do not have important consequences will have no washback.
H. 14. Tests will have washback effects on all learners and teachers.
H. 15. Tests will have washback effects for some learners and some teachers, but not for others.

As is seen, tests have more effects than considered, and they can have an impact on a wide range of factors such as teachers, learners, teaching, learning, etc.

2.2.5. Authenticity

Authenticity is defined as “the degree of correspondence of the characteristics of a given language test task to the features of a target language task”, and it “relates the test task to the domain of generalization to which we want our score interpretations to generalize” (Bachman and Palmer, 1996, pp. 23-24). Brown (2003, p. 28) expressed that a test may be authentic when it has the following features: natural language, contextualized items, meaningful topics, real-world tasks and the existence of thematic organization such as story-line.

In summary, above are the principles of language assessment. What is crucial is taking all these principles into consideration and making use of them while assessing learners. To be able to utilize them, the following questions can be asked by the teacher for increasing the effectiveness of assessment (Brown, 2003, pp. 31-37):

Are the test procedures practical?

Is the test reliable?

Does the procedure demonstrate content validity?

Is the procedure face-valid and "biased for best"?

Are the test tasks as authentic as possible?

Does the test offer beneficial washback to the learner?

2.3. Assessment of Language Skills

A number of competencies are viewed to construct L2 ability, and the competencies underlying L2 ability are reading, listening, writing, and speaking. Below includes detailed information about the assessment of these four skills under separate headings.

2.3.1. Assessing reading

Reading is regarded as a receptive skill that cannot be observed directly, and its importance is clear that people access information by reading because much of the information comes from written sources (Hubley, 2012). As the process of reading is not directly observable, while assessing reading skills, the subskills which are believed to constitute reading skills are taken into consideration (Hubley, 2012). These subskills are generally discussed under three headings such as bottom-up, top-down and integrative approach. Hubley (2012) explained these three approaches in detail. Bottom-up approach is the oldest among them, and it was put forth in the 1930s. The smallest units such as morphemes, letters are the foci, and they are decoded. Top-down approach focuses on the larger parts and bits of the reading text. Skimming in order to find the main idea is the focus, and along with main ideas, supporting sentences are sought by readers. Then, interactive approach came to the ground embracing both approaches that are bottom-up and top-down. The interactive approach posits that readers both attend to the global meaning of the text by paying attention to larger units, and to the local meaning of the text by attending to details. Additionally, the relationship of the reader with the text is also important in interactive approach in which "testing tasks may require students to

recognize how parts of the text are interconnected with discourse markers or to detect shifts in opinion that are supported by specific details" (p. 212). The stages of reading are associated with interactive approach. These stages are pre-reading, while-reading (or during-reading) and post-reading (Farrell, 2008). In pre-reading, students' schemata is activated, and they try to predict what will be the next in the lesson. This prediction is based on their previous experiences with the topic; so, topic familiarity is important in this stage. In while (during)-reading, learners use the text, and the pictures related to the text to confirm their predictions in pre-reading stage. In post-reading stage, students are required to cover and analyze the text, and they are asked comprehension questions to check their understanding of the text.

Hughes (1989) divided reading skills as macro-skills and micro-skills. Macro-skills include scanning, skimming, identifying stages and examples of an argument. Micro-skills are composed of identifying referents, using context to guess the meaning of the word, and understanding relations between parts of the text.

Knowing the subskills that make up a reading skill is crucial as stated in the previous paragraph. What is as important as this issue is how reading skills should be assessed, and more specifically, the selection of proper reading texts. Harris (1969) suggested that while selecting the reading texts, length, subject matter, style and treatment of subject, and language be taken into consideration. In terms of length, the reading passages should be kept short enough to provide necessary context for readers to comprehend them. In terms of subject matter, so as to understand the reading passages, readers should not resort to their outside knowledge. If so, resorting to outside knowledge makes certain readers more advantegeous than the others. Additionally, the issues known by everybody should not be the subject of reading passages. Concerning style and treatment of subject, various types and styles should be integrated into the reading exam. As for the language, reading texts should not be loaded with extremely difficult grammatical structures, and lexical items. Simplification can be made but more care should be given to the simplified reading text because, at that time, there is the risk of not being able to discriminate learners due to its being too easy. These issues are the factors that are also thought to affect the difficulty of reading texts by Alderson (2000) who stated that, in addition to aforementioned factors above, there is also the factor affecting the difficulty of reading texts which is the presence or absence of reading texts. It is commonly believed in the literature that removing the text just before the learnes are

answering the questions increases the memory effect which is not the purpose of assessing reading skills.

In addition to the careful selection of reading passages, another issue that should be given great care is writing items to assess reading skills. Alderson (2000) stated that there are certain factors which influence the difficulty of test items in reading. The first factor is language of questions. If the questions cannot be fully understood by learners owing to the difficult language, then it becomes difficult to say that poor performance of learners result from the reading text or difficult questions. The second factor is the type of questions. Pearson and Johnson (1978) came up with three different types of questions that could be asked to assess reading skills. These are textually-explicit, textually-implicit, and script-based questions. When the question and the answer can be found in the same sentence, these questions are called textually-explicit questions. If learners are required to combine sentences to obtain the necessary information, then they are called textually-implicit questions. The last one is the script-based questions in which learners are expected to resort to their background knowledge and combine it with the information gotten from the text in order to answer the questions because the answer is not stated in the reading text.

As an alternative to Pearson and Johnson (1978)'s discrimination, Alderson (2000) divided question types into two as global and local questions. This division is similar to the division between textually-implicit and textually-explicit questions. Harris (1969) warned that while writing items for assessing reading skills, there are some issues that should be taken into consideration. First one is that vocabulary and grammar of the items should not be too difficult for readers. Secondly, the stem of the item-the question itself, not the options- should introduce the problem, and give some clues about the question to be asked. The other one is that mere matching is not desired. Students should not be asked to match the words in questions with the same words in the reading passage. The last one is all the questions could be answered after carefully reading the text. Outside knowledge, the conflicts among the options, or illogical options should not help the learner eliminite some options.

As for the techniques for assessing reading skills, Alderson (2000) stated that there is not a method which is perfect. The commonest way of assessing reading skills is through multiple choice questions. Cloze test and gap-filling test have also become prevalent in assessing reading skills since they are easy to develop. There are some

objective tests such as matching tasks, ordering tasks, dichotomous items (true-false), and editing tests (there are some errors in the text, and learners are expected to identify them). Other techniques that are not acknowledged as objective as the previous ones are short answer, information transfer, summary test, gapped summary, free-recall test, and c-test. Apart form the techniques used to assess reading skills, Hughes (1989, pp. 126-129) stated that there are some techniques used for specific purposes that are identifying the order of events, topics, or arguments, identifying referents, and guessing the meaning of unfamiliar words from context.

For Brown (2003), the division of assessing reading based on the tasks is as follows: perceptive, selective, interactive and extensive performance. In perceptive performance, the purpose is to comprehend the parts of a larger text including letters, punctuation, etc. Bottom-up processing is utilized here. Typical tasks are reading aloud, written response tasks and multiple choice questions (limited). Second type is selective performance in which learners are expected to attend to lexical, grammatical or discourse features. Not only bottom-up processing but also top-down processing could be utilized in this, and learners are expected to produce limited responses. Picture-cued tasks, matching, true/false, multiple choice questions (for form-focused criteria), and editing tasks can be classified under this heading. Third one is interactive performance which sees reading "a process of negotiation of meaning", and "the reader brings to the text a set of schemata for understanding it, and intake is the product of that interaction" (p. 189). Typical examples can be comprised of information transfer, cloze task, editing (longer texts), ordering tasks, anecdotes, short narratives and descriptions, announcements, etc. Top-down processing is utilized mainly here, but there may be need for bottom-up processing as well. The last type is extensive performance that includes professional articles, essays, technical reports, short stories and books. The purpose is to trigger learners' main understanding, and to be able to achieve this, top-down processing is used here. Skimming tasks, summarizing and responding, note-taking and outlining can be categorized under this type.

2.3.2. Assessing listening

"Listening has often been described as the "Cinderalla" skill (Flowerdew, 1994; Nunan & Miller, 1995; Flowerdew & Miller, 2005), and it is the language skill most teachers take for granted, and skill any students spend less time on actively developing"

(Flowerdew & Miller, 2012, p. 225). For most teachers and students, listening is a skill that can improve itself, and you do not have to make many efforts for this; however, many recent research studies have demonstrated that listening is a skill which plays a crucial role in the development of other skills (Flowerdew & Miller, 2012). In the same vein, Buck (2001, p. 247) also stated that listening is a "complex process in which the listener takes the incoming data, an acoustic signal, and interprets it based on a wide variety of linguistic and nonlinguistic knowledge". This process makes comprehension an on-going process, and "meaning is actively constructed by the learner" (p. 247).

Buck (2001) came up with three approaches to assess listening skills which are the discrete-point approach, the integrative approach, and the communicative approach. The discrete-point approach is based on structuralism and behaviorism. Isolated items are tested, and they are tested independently. Phonemic discrimination tasks are typical examples of this approach. The second one is the integrative approach that tests more than one item at a time. The items are not tested separately and independently anymore in this approach. Gap-filling exercises and dictation are typical examples for the integrative approach. The last one is the communicative approach in which listeners are expected to apply what they have understood in wider contexts. Being proficient in language means "being able to demonstrate a degree of communicative competence (Hymes, 1972)" (p. 226). In this approach, real use of language is fostered, communicative peformance is emphasized rather than linguistic accuracy, activities are close to the real life examples, tests have communicative purposes and authenticity is given importance (Weir, 1990). These are the main features of communicative approach to assessing listening skills. However, Flowerdew and Miller (2012) demonstrated that there are several problems concerning the communicative approach. To begin with, it is more difficult to develop communicative tests when compared to discrete-point and integrative approach. Second one is that there is no exact and correct way to react in one situation. As there are some different possibilities of performing in one situation, it is not very easy to assess listening. The last one is that communicative events abound; as a result, testing all communicative events is nearly impossible.

Despite the fact that there are some difficulties related to the assessment of listening skills through communicative approach, there has been a growing interest and popularity of communicative approach to assessing listening in the last years. The possible reasons for this growing interest and popularity of communicative approach could be as follows.

The first one is communicative tests are contextualized and more authentic when compared to other approaches mentioned above, and the other one is that a person has the chance to encounter these situations in real life which makes these communicative tests purposeful (Flowerdew & Miller, 2012, p. 226).

In addition to the importance of approaches that could be utilized while assessing listening skills, the techniques that are used while assessing are also worth mentioning. The techniques used are crucial, because these are the way how listening skills are assessed. For Hughes (1989), there exist some possible techniques to be utilized in assessing listening skills. These are multiple choice, short answer, information transfer, note-taking and partial dictation. Buck (2001) came up with more techniques to be used under the headings of discrete-point, integrative and communicative approach. The techniques in discrete-point approach are phonemic discrimination tasks (often called minimal pairs such as wine-vine), paraphrase recognition, and response evaluation. Integrative approach consists of noise tests, listening cloze, gap-filling techniques, dictation, sentence-repetition tasks, statement evaluation, and translation. Noise tests and listening cloze are covered by the term reduced redundancy in which "elements are removed thus reducing the redundancy of the text" (p. 68). Reduced redundancy takes pragmatic expectancy grammar as its basis. In noise tests, there is background noise accompanying the passage. The last one is communicative approach which favors the use of authentic tasks and tasks which have a communicative purpose to assess listening skills.

Brown (2003) divided the assessment of listening skills into four types as intensive, responsive, selective and extensive performance, and distinguished the tasks that are used to assess listening skills under these four types of listening. Recognizing phonological and morphological elements and paraphase recognition are included in intensive performance which is listening for understanding the parts of a larger unit such as words and phonemes. Choosing appropriate response to a question and writing a response to a question are in the second type which is responsive performance. In responsive performance, learners listen to a short unit of long language, and give a short response to that. Listening cloze, information transfer, and sentence repetitition are covered by selective tasks in which the purpose is to scan information. Overall meaning is not the main focus here. The last type that is intensive performance includes dictation, communicative stimulus-response tasks, and authentic listening tasks such as note-taking,

editing, interpretive tasks and retelling. In intensive assessment of listening, learners listen to "develop a top-down, global understanding of spoken language" (p. 120).

Along with the approaches to assessing listening and the tasks used to assess listening skills, another issue to take into consideration is to whether recordings or live presentations should be utilized (Hughes, 1989). Opposite points of view exist in the literature concerning the use of recordings or live presentations; so, this issue is under discussion. When recordings are used, there is uniformity in what is presented, but there is the need for good quality of sound system here. On the other hand, live presentations are appropriate as long as the same person gives the speech. Hughes (1989) attracted attention to the fact that all the speakers giving the speech which is used for assessing listening skills should be trained and have a good command of language. If there exist a lot of classes in which students hold the exam, then it is not very appropriate to expose students to different speakers. Instead, recordings should be preferred in these situations. Buck (2001) stated that using recorded stimuli has advantages of authenticity; on the other hand, live presentations are more advantageous owing to requiring no technical equipment and being easy to administer.

Buck (2001) also gave information about the purposes for assessing listening skills. Very first of these is for general language proficiency. Listening is one of the four main skills which make up a language ability. People spend nearly half of their communication time on listening (Feyten, 1991); thus, listening should be given importance. Second one is for representing oral skills. Buck (1991) believed that listening can be used instead of other oral skills because testing speaking is time-consuming, expensive, and requires great resources. The next one is assessing listening for achievement purposes. If listening is taught, then it should be tested as well. Whether learners can have the necessary knowledge about listening skills to move on the next grade is decided based on the scores gotten from achievement tests. Additionally, if listening skills are tested, then learners are given a motivation to practise this skill. The last one is for diagnostic testing. If the weaknesses and strengths of learners could be identified, then instruction will be more effective and to the point to the learners' needs, because instruction can be adapted according to the learners' needs.

2.3.3. Assessing writing

Harris (1969) stated that writing was utilized for the purposes of reinforcing the learned grammatical structure or lexical unit in the past; but then, it was acknowledged as a separate skill, and was treated "as an end in itself- as a complex skill involving the simultaneous practice of a number of very different abilities, some of which are never fully achieved by many students, even in their native language" (p. 68). In a similar fashion, in order to draw attention to the fact that assessing writing is not a simple activity, Weigle (2012) expressed that assessing writing seems such a simple issue in which students are given a topic and they are expected to write on that topic; however, that is not the case.

Many concerns exist in the literature regarding the existence of timed writings to assess writing skills or not, and how these skills should be tested. To test or not to test writing abilities of learners is under discussion among scholars, and many reasons have been identified for their justification. One is that normally people do not write under timed conditions in real life; thus, whether timed writings in the class reflect the writing abilities of teachers leaves a question mark in minds. The other one is that teachers do not prefer to allocote all their class time to writing in class; so, they ask their learners to write their writings out of the class (Weigle, 2012).

Weigle (2012) discussed that there are certain reasons for being favorable towards asking learners to write timed writings. To begin with, teachers have the opportunity to see what and how learners can write without any help in this limited time. Also, many high-stakes exams include timed writings; hence, the timed writings teachers ask learners to write about may be the practice for the high-stakes exams. Finally, since "a writing test may function as a measure of automatized language knowledge" (p. 219), through timed writings teachers could have the chance to learn about the true picture of their learners in terms of their writing abilities. On the other hand, Weigle (2012) discussed the reasons why learners should not write timed writings. First one is that asking learners one or two topics restrict them, and it may not be appropriate to judge learners by just giving one or two topics. Secondly, non-classroom writing reflects real writing better, because learners can make use of other sources or dictionaries. Thus, the idea of portfolio can be a good option. In portfolio assessment, students give many samples of their work to the teacher, and the earlier versions of these samples have been revised by the teacher and peers for feedback.

For the assessment of writing skills, Harris (1969) expressed that the most direct way of assessing writing skills is to ask learners to write, and added that asking learners to write has been exposed to many criticisms in the literature. The proponents of this defended themselves by saying that when a learner is asked to write, certain writing abilities such as ability to organize, relate can be measured in a more efficient and detailed way than objective tests. Secondly, if students are not asked to write, then they will reject to write in the class as well, and accordingly they will not be very eager to write. The last one is composition tests are easy to develop; so, it is practical for teachers to develop them. The opponents of composition tests thought that asking learners to write is unreliable, and the scoring is very subjective. They also stated that students have a chance to avoid using certain structures while writing an essay; however, avoidance is not possible if objective tests are developed. The last one is when practicality is of concern, it is much easier to score objective tests than composition tests.

Even though there have been many problems concerning the appropriacy of making use of composition essays, there are some ways of increasing the effectiveness of them (Harris, 1969; Hughes, 1989). To begin with, learners could be asked to write several samples instead of one. Secondly, the topics that require students to have some kind of creativity and intelligence should be avoided, because the purpose is to test learners' writing abilities, not anything else. The next one is that learners should be guided by the clear instructions of the writing task. After reading the instructions, they should be clear about what is asked from them. In other words, learners should be restricted by the instructions given to them, because this limitation makes it easier for teachers to be able to compare and contrast their learners' written work. Finally, giving no options for the tasks is an important issue, because when all the learners write on the same issue, then their performance could be compared to each other. Hughes (1989) also suggested that while assessing writing skills, the demand on the learners' reading skill should be minimized, and a way of achieving this could be by making use of illustrations.

Apart from the care which should be given to the design of composition essays, great care should also be given to how these composition essays should be analyzed. Harris (1969) discussed five general components of assessing a written work, which are agreed by many scholars in the literature. These components are content, form, grammar, style, and mechanics. It is obvious that writing is not a simple issue; rather, it is difficult

because it requires many various elements making writing “a highly sophisticated skill” (p. 69).

When it comes to scoring in assessing writing skills, Hughes (1989) mentioned that there are two kinds of scoring which are holistic and analytic scoring. Holistic scoring is also called as impressionistic scoring, because overall impression accounts in this type of scoring. In analytic scoring, each aspect is assigned a separate score. In terms of being rapid -so practical-, holistic scoring has advantages over analytic scoring. However, analytic scoring has many advantages over holistic scoring in certain aspects. First of all, not all subskills underlying writing ability can develop at the same rate. Secondly, it is more reliable because there are many scores assigned to separate sections. The last one is that many aspects are taken into consideration and evaluated by the rater, which will be possibly ignored in holistic scale. After explaining the difference between these, Hughes (1989) concluded that which scale to be used depends on the purpose. If the purpose is to see the strengths and weaknesses of the learners in different subskills, then it is more beneficial to utilize analytic scoring. On the other hand, if the purpose is to see whether learners are proficient or not at the end of the term, then it is better to use holistic scoring. The number of the learners is also an important issue, because if there are many learners, then it is not practical to assess their writing skills by referring to analytic scale owing to its being detailed and taking more time.

For Harris (1969), the criticisms toward scoring of composition tests could be minimized if certain steps could be followed. To start with, decision must be made on the points allocated for certain parts of the composition. Second important issue is scoring the written work without seeing the names of the learners, because knowing the student may affect the scoring of the teachers, which in turn makes scoring more subjective and unreliable. Scanning all the papers is another way of minimizing the problems that are linked with the scoring of composition tests. The last one is having at least two raters, and getting the average of two scores as a final score.

As is clear in the literature, the most direct way of assessing writing skills is to ask learners to write. Still, there are many ways of assessing learners’ writing skills through different techniques and tasks. Brown (2003) classified the tasks designed for assessing writing skills into four, which are imitiative, intensive (controlled), responsive, and extensive performance. First one, imitiative performance, aims at assessing the skills which are called as mechanics. Primary concern is form here, not meaning. Spelling of

the words and phrases are on the spot in this type. Second one is intensive performance in which learners are expected to form the correct words in a context. Meaning and context are given importance to a certain extent, but more importance is still on the form. The next one is responsive performance that requires learners to perform at a limited discourse level. More emphasis is on meaning and context. The last one is extensive performance in which learners are required to combine all the skills that form the process of writing such as organizing and developing ideas logically, supporting ideas through examples, etc.

In terms of the principles that a writing task should have, Weigle (2012, p. 220) discussed the qualities of a good test which are reliability and validity in relation to assessing writing. A writing task can be reliable when the test situations are the same for all learners, when learners are given the same amount of time to finish the writing task, and when they are given the same topic to write about. When it comes to scoring, reliability could be increased by having at least two raters who use the same criteria to score the written work. A writing task can be valid when some of the learners are not adventageous because of their familiarity and interest in the topic which is given to them to write about, and can be valid when the topic given is the representative of all the content.

2.3.4. Assessing speaking

Valette (1977) stated that "speaking is a social skill. One can read and write in private or listen to the radio or watch television alone; however, it is rare for a person to speak without an audience of some sort. In brief, oral communication is the goal of speaking and it requires a speaker, a listener, an interaction" (p. 119). In the same vein, it is not very easy to design oral communication tasks that disregard listening totally (Brown, 2003). As is seen, it is not very easy to separate speaking from listening.

In addition to the comparison of speaking and listening, Harris (1969) also discussed the similarities and differences between speaking and writing that are both acknowledged as productive skills. Speaking is like writing in the sense that both are complex skills that learners should make use of various abilities at the same time, and these abilities do not develop at the same rate interestingly. On the other hand, speaking is not similar to writing in the sense that writing is a more formal and sophisticated skill, and there can be people who never fully master writing even in their mother tongue. For

speaking, the main concern is to be able to communicate informally in daily life, and that can be achieved easily and fluently by many people in their first language.

O'Sullivan (2012) stated that assessing speaking is believed to be the most difficult to develop and administer (p. 234). There exist several reasons for this belief. Even though there is paucity of work in assessing speaking in the literature, some research studies have dealt with assessing speaking in relation to certain reasons (O'Sullivan, 2012). These factors are related to "the impact on performance of characteristics of the test taker, of the interlocutor, of manipulating task performance conditions, of the development and use of different types of rating scales, and of rater or marker performance (p. 234).

As for the test design, O'Sullivan (2012) gave information about the common types of test design that are utilized while assessing speaking. The most common type is interviews which are mostly designed in one-to-one format. It is the easiest way of assessing speaking skills, because it is very practical. What is needed is only the teacher and the student. Second one is monologue in which learners are expected to give a short speech on a given topic after getting prepared. Learners are given some time before they get ready for their speech. It is limited in its function, because there is no interaction here; however, it is favored by some teachers because the teacher has the control over what the learners are going to say and length of the speech. The other one is small group interaction in which two or more learners are expected to discuss on a given topic during which they are assessed. The drawback of this design is more dominant students may not give chance to less outgoing learners to speak. The last one is the recorded stimuli which usually takes place in a language laboratory, and learners record their voices while answering the questions or discussing the given topic.

For Harris (1969), the division of test design in assessing speaking is as follows: relatively unstructured interviews, highly structured speech samples, and paper-and-pencil objective tests of pronunciation. For the interviews, the biggest weakness is even the same rater evaluates the interview for the second time, the scoring will be different, which in turn makes this type unreliable for some scholars. Highly structured speech samples came to the ground owing to the weakness stemming from interviews in which different tasks are given to different learners, and based on these different tasks, learners are given scores. The patterns under highly structured speech samples are sentence repetition, reading passage, sentence conversion, sentence construction, and response to

pictorial stimuli. As for the paper-and-pencil tests of pronunciation, rhyme words, words stress, and phrase stress can be grouped in here.

Brown (2003) expressed there are five kinds of speaking performance and these aforementioned test designs could be placed under these categories. These types are imitiative, intensive, responsive, interactive, and extensive (monologue) performance. In imitiative performance, learners are expected to imitate a word, phrase or a sentence, and the main emphasis is on pronunciation. Word repetition task is a typical example of this type. Intensive performance requires learners to form short sentences, and minimum interaction is involved in this type. Typical examples could be as follows: directed response tasks, reading aloud, sentence and dialogue completion, limited picture-cued tasks, and translation at the sentence level. Third type that is responsive performance includes interaction and understanding of the test. Question and answer, giving instructions and directions, paraphrasing, small talk and standard greeting can be put under this category. The next sort is interactive performance in which interaction is longer and more complicated when compared to responsive. Interviews, role-plays, discussions and conversations, and games can be considered in this category. The last one is extensive performance which is comprised of speeches, oral presentation, story-telling, retelling a story/event, and translation (of extended prose). Interaction is limited, sometimes no opportunity is given to learners for interaction.

When it comes to the criterial levels of performance, there are no best criterial levels to assess speaking skills. There exist differences in the list of scholars in the literature. The followings are the examples of this. Assessing speaking is composed of five elements that are pronunciation, grammar, vocabulary, fluency, and comprehension (Harris, 1969). Also, Hughes (1989) stated that these criterial levels include accuracy, appropriacy, range, flexibility, and size. As in the assesment of writing skills, holistic or analytic scales could be utilized to assess speaking skills. As Hughes (1989) suggested, it is better to make use of one type of scale to support the other type of scale; so, both can be used together to compansate the weaknesses of each other. What is important in both kinds of scales is the training of the raters, not evaluating learners in terms of their linguistic ability solely, and having more than one rater (Hughes, 1989).

2.4. Assessment Literacy

Assessment is an indispensable component of teaching and learning process, and the big role of teachers in assessment is undeniable. Teachers have many roles in assessment-related activities. Because assessment is used for monitoring, placement and ranking purposes, teachers have to be aware of these uses of assessment, and teachers are expected to be familiar with the concepts related to these external uses of assessment (Inbar-Lourie, 2008). As the use of assessment is not limited to these only, teachers have to be aware of the difference between assessment and testing cultures that are contradictory in terms of assessment. As a result of this contradiction, teachers have to be knowledgeable in both cultures by being familiar with alternative assessment which is assessment culture and by abiding the rules of external authorities in testing culture (Inbar-Lourie, 2008). McMillan (2000, p. 1) stated that assessment is a process including professional judgment of teachers, and professional judgments cover "constructing test questions, scoring essays, creating rubrics, grading participation, combining scores, or interpreting standardized test scores". Teachers should also know the difference between the following terms which are formative and summative, criterion-referenced and norm-referenced, traditional and alternative, standardized tests and classroom tests (McMillan, 2000). McMillan (2000) added that knowing these terms is not enough for a teacher. Knowing these and deciding on which one to use to promote instruction is important in assessment.

Fulcher (2012) stated that language testing and assessment have undergone a huge change in the first part of the 21st century, which in turn leads to the change in the needs of language teachers. This change gives more importance to the term "assessment literacy". There are three reasons for assessment literacy to be very important (Fulcher, 2012). To begin with, the use of tests and assessment has increased a lot. Second one is the increased use of tests and assessment as part of national immigrant policy. Though the first two reasons are external to the field, the last one is internal. Assessment for learning has been very popular in the field; so, assessment has become a component of classroom practice. Although these aforementioned reasons have made assessment literacy much more important than ever, the problem is that what constitutes assessment literacy exactly still remains a question (Fulcher, 2012). There are many scholars in the literature trying to define this term and what it covers. Some of them are as follows:

Assessment literacy is seen as a bridge between learner achievement and the quality of assessment (Mertler, 2009). Davies (2008, p. 328) defined it as the combination of skills and knowledge. Skills refer to knowing how to construct and analyze a test, and knowledge refers to the "relevant background in measurement and language description". For Falsgraf (2005, p. 6), assessment literacy is "the ability to understand, analyze and apply information on student performance to improve instruction". Thus, being an assessment literate requires some properties such as having theoretical and practical competencies and also knowing how and why to construct a variety of assessment procedures (Boyles, 2005). For Popham (2004), assessment literacy is the understanding of sound assessment, and for Stiggins (2007, p.2) knowing the difference between sound and unsound assessment. Xu and Brown (2017) stated that "assessment literacy is central to the quality of education because competencies in assessing student learning lead to informed decisions" (p. 133), and assessment literacy is considered as a part of teacher expertise (Xu & Brown, 2016). "Understanding teachers' current levels of assessment literacy mastery is a good departure point for promoting both assessment literacy research and teacher development in education" (Xu & Brown, 2017, p. 134). It is obvious from the definitions above, there are many scholars who attempted to define assessment literacy in the literature. Regarding this situation, Coombe (2012, p. 2) stated, "the definitions of the term assessment literacy abound in the literature".

Though the definitions are many in number, the ideas are the same across all of the definitions. Teachers are in the center of assessment, and they are expected to possess certain skills to carry out assessment-related activities effectively.

The American Federation of Teachers, the National Council on Measurement in Education, and the National Education Association (1990, pp. 31-32) came up with seven standards that are crucial for Teacher Competence in the Educational Assessment of Students. These standards are accepted as "an important landmark in defining teachers' assessment literacy" (Inbar-Lourie, 2017, p. 259). The standards are as follows:

1. choosing assessment methods appropriate for instructional decisions;
2. developing assessment methods appropriate for instructional decisions;
3. administering, scoring, and interpreting the result of both externally-produced and teacher-produced assessment methods;
4. using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement;
5. developing valid pupil grading procedures which use pupil assessments;

6. communicating assessment results to students, parents, other lay audiences, and other educators;
7. recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information.

Assesment literacy is seen as a "sine qua non for today's competent educator" (Popham, 2009, p. 4). In his paper, Popham (2009) touched upon assessment literacy by comparing two ideas: whether it is a must for educators, or it is like a fashion to be forgotten soon. He insisted on the fact that assessment literacy is a must, and what is implied in the term assessment literacy is teacher's knowing the assessment methods used in the classes. In other words, it is the familiarity of the teachers with assessment-related terminology. He added that assessment literacy should be the focus of professional development programmes till all pre-service education programmes produce assessment literate teachers. Thus, what can be concluded here is that pre-service education does not equip pre-service teachers with the necessary and adequate assessment knowledge to be assessment literate teachers in their profession.

Teachers' being assessment literate is crucial, and in the same vein, if they do not have enough assessment literacy then it poses a problem because "insufficient assessment literacy leads to reduced reliability and validity, and further results in mis-directed and ill-formed decisions" (Xu & Brown, 2017, p. 134). Furthermore, as the teachers' not having adequate knowledge in assessment can "cripple the quality of education" (Popham, 2009, p. 4).

Mertler (2009) expressed that teachers as assessors should be aware of the wide range of assessment-related activities and their strengths and weaknesses, and this awareness is a prerequisite for good assessment. Teachers should also have the ability to adapt their paradigm to comprehend the impact of assessment on learning and the performance of learners, because assessment can drive instruction (Davidheiser, 2013). As is clear, teachers have various roles as assessors, and in order to carry out all these assessment-related activities properly and effectively, a teacher must have assessment literacy. Despite the fact that teachers should be surrounded by sound assessment knowledge and practices, Popham (2009, p. 5) drew attention to the lack of assessment knowledge of teachers by saying that "test is a four-letter word, both literally and figuratively". Though sound assessment requires assessment literate teachers, teachers do not have necessary knowledge and skill in educational assessment, unfortunately.

Stiggins (1991) stated that many learners graduating from most of the educational programmes are not confident and knowledgeable enough to assess their learners in school context, because they are not prepared enough in how to assess their learners. Similarly, many teachers in the US reported that they do not feel themselves equipped with the necessary assessment knowledge in order to assess learners, and they believed that the assessment training did not make them prepared to carry out assessment-related practices effectively in teaching process (Mertler & Campbell, 2005). Popham (2004) also drew attention to the inadequate assessment literacy of teachers. Mendoza and Arandia (2009) stated that pre-service teachers needed more training to be better at assessment practices, and in order to achieve this, teacher education programmes in higher education have many responsibilities to train language teachers in assessment.

As is obvious from the statements above, the importance of assessment literacy for teachers is undeniable, and it is, for sure, not an extra feature a teacher needs to possess, rather it is a must for all teachers. Coombe (2012, p. 2) discussed the importance of assessment literacy for teachers under separate paragraphs, both from a practical and empirical point of view. What is expressed can be considered as a summary of the literature about why assessment literacy is crucial for teachers. First one is the most important one for Coombe, stating that when English language teachers have a solid background, they are well-positioned to combine and relate assessment with teaching. When they can have this background, they can achieve to differentiate the purposes of assessment and make use of them properly. Secondly, as Stiggins (1995) uttered teachers are engaged with assessment in half of their time which cannot be disregarded for teachers. Owing to this, it is clear that it is not an extra thing for teachers to know about sound assessment. Then, they have to be aware of assessment and assessment-related activities. Third one is related to the relationship between assessment literacy and professional development for teachers. When teachers are aware of sound assessment and why they utilize them, it is more probable to talk about increased test validity and the promotion of transparency, which will result in the communication to all stakeholders more effectively in the end. The last one is related to student achievement. When teachers can make use of assessment-related activities effectively, it will directly have a positive impact on student achievement, which can be regarded as an empirical perspective.

Webb (2002, pp. 1-2) discussed the need for teachers to be assessment literate under two reasons. One is the advent of standards-based reform which has resulted in more

explicit learner expectations and accordingly the immediate need to determine if learners have realized these expectations or not. The other one is the great acceptance of utilizing various forms of assessment such as norm-referenced and criterion-referenced. Webb (2002) made it obvious that there are many ways of assessing learning containing standardized tests, performance tests, and portfolios. An assessment literate teacher is capable of all these types of assessment, why they are used, and how they are used. Additionally, the necessity for teachers to have sufficient assesssment knowledge was also made clear in the sentences of Popham (2006, p. 85) who said that "today, more than ever, assessment plays a pivotal role in the education of the students. That's why educators –and everyone else who has an interest in education- need a dose of assessment literacy".

As is clear in the literature, there is agreement among the scholars that assessment literacy is not a luxury for a teacher, rather, it is a must every teacher has to possess. As a last word but not the least, in order to show how crucial assessment literacy is for teachers the following can be used as a kind of summary: "without a higher level of teacher assessment literacy, we will be unable to help students attain higher levels of academic achievement" (Coombe, Troudi, & Al-Hamly, 2012, p. 20). As the aim is to support learners in their learning process in many aspects, the key to this support is being more assessment literate in terms of teachers.

2.5. Teachers' Assessment Knowledge

The main goal of assessment is to "support and improve both learning and teaching; therefore, it is imperative that teachers examine their knowledge, practices, and beliefs in relation to language assessment" (Haught & Crusan, 2016, p. 179). Xu and Liu (2009) found out that teacher's prior experience of assessment, power-laden relationships, and specific location play important roles in shaping teacher's assessment knowledge construction. As an important concept in the assessment process, accordingly in teaching and learning process, assessment knowledge is regarded as a main component of the knowledge base of teachers more and more nowadays (Stoynoff & Coombe, 2012). DiRanna et al. (2008) also discussed the importance of assessment knowledge by stating that teachers should combine their knowledge of teaching and knowledge of assessment in order to be more effective in their instructional decisions. By drawing attention to the strong link between assessment knowledge and assessment literacy of teachers, Xu and

Brown (2017, p. 134) stated that assessment literacy should start with the investigation of its knowledge base; thus, the assessment knowledge is the heart of assessment literacy.

However, what Popham (2009) uttered did not draw an inspiring picture for the assessment knowledge of teachers by displaying that a vast majority of teachers do not know much about educational assessment, and for some teachers "test is a four-letter word, both literally and figuratively" (p. 9) which is understandible because most of the teachers' exposure to the concepts and terminology in assessment is limited to their a few class hours in pre-service education.

As a few class hours or even a course throughout a term cannot be sufficient for such an important issue, Stoynoff and Coombe (2012) discussed the relationship between professional development and language assessment literate teachers. Professional development is a must for teachers, because they are "expected to choose or construct, administer and interpret the results of assessment designed for a variety of purposes and situations" (p. 122). They added that the reason why teachers should be more assessment literate, that is why they start professional development or continue it, may differ, but "developments such as the establishments of standards by professional groups, the implementation of government policies, and the introduction of educational change are some of the factors prompting teachers to pursue professional development in language assessment" (p. 122). In parallel with the aforementioned statements, Popham (2006) also expressed that there is a need for an ongoing in-service assessment training which is in parallel with the assessment practices.

2.6. Studies on Assessment Literacy and Assessment Knowledge of Teachers

Though the importance of assessment and assessment literate teachers have been made clear in the literature, the change towards more assessment literate teachers has been slow (Coombe, Troudi, & Al-Hamly, 2012). As is clear in the literature that teachers do not have adequate assessment literacy. Being aware of the need to measure assessment literacy levels of the teachers in order to detect the strengths and weaknesses of teachers, Impara, Plake and Fager (1993) developed Teacher Assessment Literacy Questionnaire (TALQ) which is a 35-item survey consisting of multiple choice questions with four options. This questionnaire was developed based on the "Standards for Teacher Competence in Educational Assessment of Students" (AFT, NCME, & NEA, 1990). The study was carried out with 555 in-service elementary and secondary school teachers in

the U.S. According to the results, the mean score of the participants was 23.2 out of 35, which was considered as low assessment literacy.

Campbell, Murphy, and Holt (2002) used TALQ with 220 pre-service teachers who had completed a course on educational assessment. The findings of the renamed Assessment Literacy Inventory (ALI) indicated that the mean score of the pre-service teachers was 21, which made it clear that pre-service teachers did not have adequate assessment knowledge. When compared with the study of Impara, Plake, and Fager (1993) who used the same instrument with in-service teachers, the findings of this study showed that pre-service teachers had less assessment literacy than in-service teachers. Mertler (2003) also studied the assessment literacy levels of teachers through utilizing an inventory called Classroom Assessment Literacy Inventory (CALI) which is an adapted version of TALQ. In this study, 67 pre-service and 197 in-service teachers took part, and their mean scores were compared via statistical analysis. It was shown that pre-service teachers answered nearly 19 questions correctly out of 35 questions. For in-service teachers, the number of the questions that were answered correctly was 22; thus, it was clear that in-service teachers did better than pre-service teachers. The results gotten from this study were similar to the findings of Impara, Plake, and Fager (1993) and Campbell, Murphy, and Holt (2002)'s studies.

Mertler and Campbell (2005) developed another instrument to measure the assessment literacy of teachers, which was named as Assessment Literacy Inventory (ALI). This instrument, having 35 items, was aligned with the "Standards for Teacher Competence in Educational Assessment of Students" (AFT, NCME, & NEA, 1990) as well. 35 questions were divided into five scenarios with each scanario followed by seven questions. A first-stage pilot test was carried out with 152 pre-service teachers, and the participants were 249 pre-service teachers for the second-stage pilot test. It was demonstrated that the mean score of the respondents was 23.83 out of 35, indicating that the pre-service teachers in this study had a low level of assessment literacy.

As discussed above, assessment literacy of teachers has been investigated by the use of some instruments which have been specifically developed to measure the assessment literacy of teachers. Some of the instruments that have been widely used in the literature are Teacher Assessment Literacy Questionnaire (Impara, Plake, & Fager, 1993), Classroom Assessment Literacy Inventory (CALI) (Mertler, 2003), and Assessment Literacy Inventory (ALI) (Mertler & Campbell, 2005).

In addition to the studies which aimed to find out assessment literacy of teachers through the instruments they developed, there are some other research studies in the field that seek to learn more about assessment literacy of teachers by means of different methodologies along with the aforementioned instruments.

Volante and Fazio (2007) carried out a study with 69 pre-service teachers from each of the four years in ELT programme. 12 of them were male, and their ages ranged from 19 to 51. The participants were given a survey including four closed and five open-ended questions. The questions in the survey were divided into four major areas such as self-described level of assessment literacy, main purposes of assessment, utilization of various assessment methods and need for further training, and suggested methods for promoting assessment literacy at university. The findings indicated that self-efficacy ratings of the participants were very low across each of the four years of the programme. The majority of the respondents made use of assessment for mainly traditional summative purposes. Furthermore, the pre-service teachers stated for an urgent need for a specific course based on classroom assessment, and this need was verbalized by all the participants across four years.

Davidheiser (2013) carried out a study with 102 teachers from various fields in education (English, Social Sciences, Maths and Science) to find out the assessment literacy levels of the participant teachers via Assessment Literacy Inventory. Three high schools which are East, South, and West of the Central Bucks School District in the USA were involved in this study. Including the interviews with four teachers as well, this study is both a quantitative and qualitative in nature. The teachers whose core-subject area was Math had the highest mean score in the questionnaire. The highest mean score was Standard 7, and the lowest was Standard 2. There was a statistically significant difference between Maths teachers and Social Science teachers, and Maths teachers and English teachers. Three themes were formed based on the information obtained from the interviews, which are assessment assumptions, assessment targets, and professional development. The participants had diverse assumptions related to assessment, and the lack of professional development was obvious.

Mertler (2009) investigated the impact of a two-week workshop for in-service teachers. Standards for Teacher Competence in the Educational Assessment of Students were the foci of the workshop, and the participants were pre and posttested by using the Assessment Literacy Inventory. Reflective journals were also utilized to obtain in-depth

information about the participants' experiences. The findings showed that the participants' performance on the posttest (M=28.89) was higher than the pretest (M=19.57). Thus, it was concluded that training had a positive effect on the assessment literacy of teachers. Besides, reflective journals showed that they had a positive attitude towards the development of assessment literacy.

The studies above were carried out with teachers who were from various fields and general education. It is natural that assessment literacy has been researched in general education and psychology more than other fields, because the term assessment literacy was rooted in these fields. Though less in number, there are some other studies trying to find out the assessment literacy levels of teachers in ELT; so, the participants in these studies are teachers whose major is English. The followings are examples of these studies in which the participants are EFL teachers.

Fulcher (2012) came up with a survey in order to detect assessment training needs of language teachers. The survey was piloted with 24 international language teachers. Language teachers were the intended participants of this study, but the ones who wanted to take part in the study could also participate in the survey; thus, the sampling became self-selecting. 278 participants responded to the survey, and both quantitative and qualitative analyses were run in order to analyze the items in the survey. The results showed that the participants were really aware of various assessment needs, and they were sure that principles and practices of assessment should be handled in a wider historical and social context. Furthermore, it was revealed that large-scale and classroom assessment should be utilized in a balanced way. As a result of this study, Fulcher (2012, p. 125) expanded the definition of assessment literacy as follows:

> "The knowledge, skills and abilities required to design, develop, maintain or evaluate, large-scale standardized and/or classroom based tests, familiarity with test processes, and awareness of principles and concepts that guide and underpin practice, including ethics and codes of practice. The ability to place knowledge, skills, processes, principles and concepts within wider historical, social, political and philosophical frameworks in order to understand why practices have arisen as they have, and to evaluate the role and impact of testing on society, institutions, and individuals".

Another study belongs to Tao (2014) who developed four different scales in order to collect data for the study. The participants were 108 EFL in-service teachers in Cambodia. These four scales are classroom assessment knowledge, innovative methods, grading bias, and quality procedure. The first one is a multiple choice test, and it aimed

to measure the assessment knowledge of teachers. In the last three scales, the aim was to find out the participants' beliefs related to assessment. All these four scales had satisfactory measurement features. Along with the quantitiave data, the researcher also made use of qualitative data to gather data. Semi-structured interviews were carried out with six teachers. The results revealed that the teachers had limited assessment knowledge, which in turn, had a negative impact on their assessment practices.

Being aware of the relationship among beliefs, knowledge, and practice, Chan (2016) explored the beliefs of 520 elementary school EFL teachers from Northern Taiwan in relation to their use of multiple assessment. The data were collected via self-report Likert scale, multiple choice and open-ended questions. Whether the participants' use of multiple assessment changes based on EFL teaching experience was also investigated in the light of the separate research question. The results displayed that the respondents had a clear understanding of what multiple assessment is and what it covers. They also believed in the effectiveness of multiple assessment, especially the use of portfolio. It was also indicated that most of the teachers tended to use more formative assessment than traditional pen and paper tests. They stated that they wanted to use formative assessment or the combination of formative and summative assessment, but none of them favored for the use of traditional assessment as the main assessment type. Besides, the results demonstrated that the relationship between the experience of the participants and their beliefs related to assessment was significant.

Sellan (2017) also aimed to get a deep understanding of the teachers' viewpoints regarding their assessment practices. The participants were English teachers in Singapore which has a distinctive Integrated Programmes (IP) context. IP does not give priority to exams, and what is important here is encouraging teacher-based assessment practices. The participants were eight teachers, and they had ten years of experience. Main data collection tool was interviews, and the researcher made use of stimulated recalls, observations and analysis of the documents as well. The findings indicated that the participants improved their assessment literacy by paying attention to culture, building on an extended understanding of genres, giving increased importance to content knowledge, and focusing on higher order skills. It was clear that the implementation of IP encouraged them to be more aware of assessment needs of the learners, and to become more assessment literate teachers.

Xu and Brown (2017) also carried out a study to get informed about the assessment literacy of 891 English teachers working in China. Adapted version of the Teacher Assessment Literacy Questionnaire was utilized to obtain the data, and this version includes 24 questions. The results showed that the most of the participants either had a basic or minimal level of assessment literacy, and the items in the questionnaire which were considered as difficult after the psychometric analysis was done could not be answered correctly by a vast majority of the respondents. It was also shown that the teachers' demographic features such as age, years of experience, qualification, title, region, assessment training experience had no significant effect on assessment literacy of teachers.

All the research studies conducted above demonstrate that the participants, including both pre-service and in-service teachers, do not have the necessary skills to be called as assessment literate teachers. Stiggins (1995) made a search on the possible factors that lead teachers to be assessment illiterate, and came up with a conclusion that there exist certain barriers to assessment literacy. The first barrier to assessment literacy is fear. The teachers feel negative emotions when they think of assessment. They have some negative connotations for the word assessment. It was revealed that the reason of their fear goes back to earlier experiences of the teachers who experienced assessment as students. What they felt as students in relation to assessment has a negative influence on their perceptions as teachers. The second barrier was aforementioned in the work of Alderson (2001). Teachers do not have the willingness to increase the level of their assessment literacy because of the fact that assessment is regarded as an extra quality for an average classroom teacher. Their perceptions displayed that teachers thought not all teachers should possess sufficient knowledge about assessment. Another barrier is related to the conditions and shared duties in the workplace. In some workplaces, there are certain teachers who are engaged with assessment and assessment-related activities, and there is no need for the others to worry about what is going on in their workplaces concerning assessment. As all the things related to assessment are given to them in a complete format, they do not have to increase their assessment literacy levels. They even do not feel pressured to increase their levels. The last barrier is concerning the resources. The teachers expressed that the resources related to assessment are insufficient, and even though administrators say that they support teachers' assessment literacy development, they do not allocate them sufficient resources and time. Administrators also think that

assessment is a natural part of a teacher's duty, and do not regard it necessary to reduce the workload of teachers in other areas to back up teachers' development in assessment. As a result, all of these negative factors come together, and cause teachers to shield against assessment literacy, and any kind of professional development to be more assessment literate teachers.

Despite the barriers mentioned above, Stiggins (2007) expressed that teachers are required to increase their assessment literacy levels in the future. In the past they were not expected to be more assessment literate, but changing times have influenced the point of view towards teachers and there is greater pressure on them to be much more assessment literate.

It is seen that not many studies exist in the literature concerning assessment literacy of teachers, especially EFL teachers. When it comes to the studies which have been carried out in Turkish context, the number of the studies gets fewer. Some of the studies conducted in Turkish context with the teachers from various fields in education are as follows:

An example for the studies whose participants include not only teachers whose major is English but also the teachers whose majors are various such as science, maths belongs to Karaman and Şahin (2014) who carried out a study with the fourth grade pre-service teachers at the Education Faculty at a state university in Turkey. Learners from seven different majors were involved in the study, including the learners whose major was English Language Teaching. In the first phase of the study, assessment literacy levels of the participants were investigated through the implementation of Assessment Literacy Inventory that was developed by Mertler and Campbell (2005). Second phase only included learners from Science Teaching. These learners were at the third grade, and they were taking classroom assessment course. They were required to prepare two lesson plans before and after this course called micro-teaching. This phase of the study was composed of both quantitative and qualitative data. The findings revealed that assessment literacy level of the fourth grade pre-service teachers was limited, and the learners whose major was primary school teaching did significantly better than the other learners. Second phase of the study displayed that after micro-teaching, there were observable improvements in the perceptions and practices of the participants regarding assessment, and this course affected their thoughts positively regarding assessment practices.

The other example aimed to translate Mertler and Campbell (2005)'s inventory into Turkish and adapt it based on Turkish context. This study was conducted by Bütüner, Yiğit, and Çimer (2010) with 260 pre-service teachers. The items in the original instrument were adapted according to the Ministry's Assessment Standards. The results yielded that overall instrument reliability was 0.859, and the psychometric qualities of the inventory strongly supported its use as an acceptable measure.

2.7. Language Assessment Literacy

Language assessment literacy (LAL) as a distinct area is rooted in the term assessment literacy (Stiggins, 1991; Inbar-Lourie, 2017). The term is highly novel, but it is drawing attention in the literature day by day. For Taylor (2013, p. 405), LAL is "potentially subordinate or overlapping category" to assessment literacy (AL). LAL has many layers and stages, and teachers are expected to have very basic assessment understanding along with having a critical eye on the assessment in these progressive stages (Taylor, 2013). Inbar-Lourie (2017) stated that the term LAL stems from AL, but LAL is different from AL in the sense that LAL "attemps to set itself apart as a knowledge base that incorporates unique aspects inherent in theorizing and assessing language-related performance" (p. 259).

Inbar-Lourie (2017) stated that LAL requires additional competencies when compared to assessment literacy, and added that LAL is the combination of assessment literacy skills and language specific skills. There are many attempts to define what constitutes LAL in the literature; so, there exist many definitions of LAL (Inbar-Lourie, 2017). Lam (2015) defined it as "teachers' understandings and mastery of assessment concepts, measurement knowledge, test construction skills, principles about test impact, and assessment procedures which can influence significant educational decisions within a wider social context" (p. 172).

Language assessment literacy was also defined as "the level of knowledge, skills, and understanding of assessment principles and practice that is increasingly required by other test stakeholder groups, depending on their needs and context" (Taylor, 2009, p. 24). Inbar-Lourie (2013) stated that LAL refers to a knowledge base, a set of competencies, or both. Brindley (2001, cited in Inbar-Lourie, 2017) came up with a framework defining LAL construct. Brindley suggested that this framework consisted of core and optional modules, and this framework was specifically designed for language

teachers. The first module was a core one, and dealt with assessment from social, educational and political perspectives. Second core module tried to define and describe proficiency by relating language assessment to language knowledge. There existed three other modules that were optional. The first two optional modules were more interested in assessment and language tests in classroom context. Finally, the last optional module "presented a more advanced discussion of language assessment and research intended for teachers planning test construction projects or assesment-related research" (p. 261).

There exist some variations in the definitions of LAL. The dilemmas mentioned above were verbalized by Inbar-Lourie (2017, p. 266) as follows: "Since the conceptualization of LAL is still in its infantile stage it suffers from growing pains, the most notable of which is an identity dilemma". Along with the various definitions existing in the literature, there is still debate about who needs language assessment literacy. The target groups differ such as teachers, testing experts, and administrators. Among these groups, the primary target group needing LAL is for sure language teachers (Inbar-Lourie, 2017). Teachers are viewed as both consumers of testing information and independent assessors (Inbar-Lourie, 2017, p. 259); hence, they have to possess "the knowledge of means for assessing what students know and can do, how to interpret the results from these assessments, and how to apply these results to improve student learning and program effectiveness" (Webb, 2002, p. 1).

Stoynoff and Coombe (2012) expressed that there are many factors causing teachers to be in need for the development of language assessment literacy, basically language assessment knowledge. To start with, the content of language assessment books has been changed in the last years. Nowadays, theory and practice are hand in hand in textbooks, and teachers are expected to develop and use assessment. Secondly, according to the results of a study conducted in 1990s, nearly half of the participants in that study reported that they had not taken a course in language testing (Bachman, 2000). This situation has gotten better in recent years by giving a chance to half of the pre-service teachers to take a standalone course in pre-service education. Thus, half of the programmes offer separate assessment courses to their students nowadays (Stoynoff, 2007). The third one is related to the new perceptions concerning language assessment that is acknowledged to adopt a cognitive and social-constructivist view. Shepard (2000) atrracted attention to the fact that what was believed in the past divides assessment from instruction, and they were seen as separate issues. However, "if a new perspective of assessment is to be fully

realized, language teachers will need to consider how their current beliefs, knowledge, and skills affect their assessment practices, and they will need to stay abreast of development in the assessment knowledge base" (p. 124).

Though there is a need and call for making use of assessment for fostering effective learning, many language teachers are not prepared to do so (Lam, 2015). The primary target population needing LAL is language teachers, but they do not feel themselves competent enough.

LAL is crucial for language teachers. Scarino (2013) stated that LAL is a necessity for language teachers, because through LAL, language teachers can "explore and evaluate their own preconceptions, understand the interpretive nature of the phenomenon of assessment and become increasingly aware of their own dynamic framework of knowledge, understanding, practices and values, which shape their conceptualizations, interpretations, judgments and decisions in assessment and their students' second language learning. Through these processes, they will gradually develop self-awareness as assessors, an integral part of their language assessment literacy (p. 311).

However, the problem is that in spite of the vital role of assessment in teaching and learning process, assessment training of language teachers are not adequate (Lam, 2015), which in turn leads to the saying of Stiggins (1991, p. 535) "we are a nation of assessment illiterates". In the same vein, Popham (2004) also drew attention to the importance of training in assessment and stated that it is not adequate; thus, it is a "professional suicide" (p. 82). As teachers are responsible for administering different types of assessment practices, assessment illiterate teachers have difficulty fulfilling in designing sound and effective assessments, and "jeopardize learning and teaching with direct consequences for students' future learning" (Lee, 2017, p. 147).

Despite the importance of language assesment literacy, Fulcher (2012, p. 117) stated that "research into language assessment literacy is in its infancy". Such an important issue has not been searched well and much in the literature. In a similar vein, Inbar-Lourie (2017) stated that except for the teacher standards assessment framework (1990), there is no document which attempts to define the particular knowledge language teachers need to possess. There may be two reasons for this situation. First one results from the fact that there is scarcity of research on language teachers' LAL needs, and second one may be related to uncertainties in the field.

As is clear, it is agreed in the literature that language teachers need to possess LAL which is indispensable part of their professions; however, what they need in terms of specific skills is still blurred. As mentioned above, there is paucity of resarch in LAL. Moreover, the instruments used to measure LAL are in the form of self-report questionnaires unlike assessment literacy inventories. Even though the instruments are different in LAL and AL, the findings were similar indicating that the teachers did not have necessary LAL, and they did not have sufficient training regarding LAL (Inbar-Lourie, 2017).

2.8. Studies on Language Assessment Literacy and Language Assessment Knowledge of Teachers

Many studies in LAL have focused on the needs of language teachers (Inbar-Lourie, 2008; Fulcher, 2012; Malone, 2013; Scarino, 2013). A special issue of Language Testing (2013) was dedicated to Language Assessment Literacy; so, this issue has contributed a lot to the understanding of LAL. Five research studies (Scarino, 2013; Malone, 2013; Jeong, 2013; O'Loughlin, 2013; Pill & Harding, 2013) appeared in this special issue, and Taylor (2013) also wrote a concluding remark in this. Below are some of the papers appearing in the issue.

To begin with, Scarino (2013) stated that teachers have both instructional and evaluative roles, and assessment literacy is a must for them. It was stated that teachers' assessment knowledge acquisition process is based on teachers' beliefs, practices, and local contexts. What is crucial is that teachers should be encouraged to form their own understanding of language assessment literacy. Second one belongs to Malone (2013) who aimed to develop the LAL of language teachers via an online tutorial programme. This online tutorial included scenarios, downloadable materials, and photographs to make testing concepts clearer, and the final form of this tutorial is called "Understanding Assessment: A Guide For Foreign Language Educators" (www.cal.org/flad/tutorial/). There were two groups in this study that were language testing experts and language teachers. To obtain data from these groups, group interviews and surveys were utilized. The results indicated that there was a difference between the reactions of both groups to the structure and content of this online tutorial. While the focus was on the expansion of knowledge of the theoretical underpinnings of the field for language testing experts, language teachers were in need of more how-to components, that is, assessment tasks.

This difference between these groups led to the discussion of the term LAL in terms of the balance between theory and practice.

Jeong (2013) also investigated feedback of the course teachers on course contents. She aimed to find out whether there was a difference between LAL levels of the teachers who were language testers and non-language testers. Language testers were defined as "individuals or professionals whose primary research interest is in areas of language testing"; on the other hand, non-language-testers were the ones "whose primary interest is in other areas of language teaching (e.g. second language acquisition) but who have had experience in language assessment-related activities" (p. 348). In total, there were 140 participants who filled in the online survey, and follow-up interviews were carried out with 13 of them. The purpose here was to investigate if certain required background is called for assessment literacy or not. The findings showed that there existed significant differences between these two groups in six areas that were test specification, test theory, basic statistics, classroom assessment, rubric development, and test accommodation. It was also demonstrated that non-language-testers felt themselves less confident in teaching technical assessment skills, and they had an inclination for more classroom assessment issues.

Finally, O'Loughlin (2013) analyzed the needs of university administrators' assessment needs because these administrators were responsible for admission decisions. These administrators were from two large metropolian Australian universities in which more than 25% of the learners were international. Learners have to take IELTS for admission to these universities, and in this study the administrators (or the researcher called them as IELTS score users) were administered a survey including questions related to IELTS use, evaluation, etc. It was concluded that administrators needed to be more assessment literate, and they needed to be educated for the valid and reliable interpretation of test scores. Then, as they were responsible for admission decisions, they would be more able to carry out these decisions with a better understanding of language assessment. Apart from the studies appearing in the special issue mentioned above, there are some other studies in the literature which investigated LAL. Some of these studies are as follows:

Beverly, Tsushima, and Wang (2014) aimed to determine the stakeholders' specific LAL needs and then came up with materials to meet these needs as the first part of a large project. There are two research questions addressed in this study, which are (1), what is

the LAL needed for users of language test scores in admission decision-making at post-secondary institutions in Canada? and (2) what useful materials can be created to develop this LAL for these score users?. Purposive sampling method was utilized, and the results of the survey were used to design workshop based on the needs of the participants. The findings revealed that the participants were really aware of the importance of LAL, and they had the willingness to develop their LAL. Al-Nouh, Taqi, and Abdul-Kareem (2014) also investigated the female EFL primary school teachers' attitudes, knowledge and skills in alternative assessment. 335 teachers were asked to fill out a survey, which is a five-point Likert scale. It has three sections that are demographic information, teachers' skills and knowledge in alternative assessment, and teachers' attitudes. It is a self-report questionnaire consisting of items with *I know how to* or *I can assess*. The follow-up focus-group interviews were also conducted with principals, head teachers and teachers. The results demonstrated that teachers' attitudes towards alternative assessment were at a medium level, and they were not very motivated to utilize alternative assessment. The teachers perceived themselves knowledgeable and skilled in alternative assessment, but some of them stated that they needed for workshops and training to be better at alternative assessment.

In addition to these, Lam (2015) carried out a study to investigate the overall language assessment training in five Hong Kong institutions, and more specifically aimed to find out how two language assessment courses facilitated or inhibited the language assessment literacy of pre-service teachers. The researcher went over ELT-related programmes based on certain criteria, and then decided upon five of these programmes for detailed analysis. In addition to gathering documents related to these five programmes such as curriculum, outline, handbook, the researcher made use of focus group interviews with 40 learners and one-on-one interviews with nine teachers from two assessment courses. All the interviews were based on getting the opinions of the participants regarding the design, content, quality and usefulness of the assessment courses in relation to LAL. The analysis of the programmes showed that there was insufficient support to foster LAL, and the training for LAL was inadequate. Based on the perceptions of the participants, three themes came out which were perceptions of LAL in an examination-oriented culture, experience of course-based language assessment training, and restricted application of LAL in authentic school contexts.

Next, Tsagari and Vogt (2017) carried out a mixed-design study covering both quantitative and qualitative data. However, in this study, they discussed the findings of the data obtained through qualitative data, namely, semi-structured interviews as a part of a bigger study. The aim was to find out the teachers' perceptions of LAL and their individual needs related to language testing and assessment. The participants were regular teachers from Cyprus (n=16), Greece (n=22) and Germany (n=25). Regular teachers in this study were defined as "the teachers who have undergone standard training and who teach foreign languages at state tertiary institutions, colleges, and schools, and have no additional assessment roles" (p. 44). The results demonstrated that the participants teachers' perceived LAL was not sufficient, and they did not feel themselves prepared effectively for assessment-related practices. Additionally, it was found that teacher education programmes were not giving the efficient and sufficient education and training in language assessment to the pre-service teachers; as a result, these programmes were not enough to prepare the pre-service learners for their future careers. Finally, the tendency towards test was dominant in most of the teachers, which in turn formed a kind of resistance in the teachers toward innovative assessment practices.

Another study belongs to Baker and Riches (2017) in which the LAL development of teachers was examined. 120 Haitian high school teachers participated in the study, and the data were collected via feedback on drafts of revised exams, survey with teachers, and teacher interviews. Some workshops were designed in 2013 for the participants, and this study took these workshops as its basis. It was concluded that LAL development of the teachers was clear after these workshops, and the main areas where the teachers' LAL levels increased were: creating reading comprehension questions, integrating vocabulary task, basing all exam sections on the same topic, increased attention of the connection between teaching and assessment, broadening of the teachers' understanding of the construct of language ability, teachers' beliefs concerning their supportive role, and finally learning about reliability, validity, and practicality.

Finally, very recently, Kremmel and Harding (forthcoming) developed an instrument called Language Assessment Literacy Survey. They had been developing it since 2015. This instrument was developed as a part of a larger project that aimed to create a comprehensive understanding of LAL that could be utilized for needs analysis, self-assessment, reflective practice and research. The instrument has recently been released in their official website; thus, there is not much information about it. There are

71 items in the survey, and it is a Likert scale consisting of five answers, from 0-not knowledgeable to 4-extremely knowledgeable. The question is "how knowledgeable do people in your chosen group/profession need to be in each aspect of language assessment?". The aspects mentioned in the question are the items of the survey some of which are identifying assessment bias, selecting appropriate items or tasks for a particular assessment purpose, and using statistics to analyse overall scores on a particular assessment. After this knowledge part is over, testtakers are presented the same items, but with a different purpose. The question is "how skilled do people in your chosen group/profession need to be in each aspect of language assessment?". The answers vary depending on the purpose of the question, from 0-not skilled to 4-extremely skilled.

Aforementioned studies were the ones that were conducted related to LAL which is a very novel research area. Due to this, the studies on LAL are very rare in number, and this number gets lower and lower in Turkish context. Hence, there are very few studies in which language assessment literacy or assessment knowledge of teachers were the foci. Some of them are as follows: To begin with, Öz (2014) aimed to investigate the perceptions and practices of Turkish EFL teachers towards formative assessment. 120 teachers took part in this study, and they were required to complete online self-report Assessment for Learning Questionnaire for Teachers which is a Likert scale. The results indicated that the teachers heavily relied on traditional methods, more than formative assessment. They also differed in their perceptions and practices related to formative assessment, and it was revealed that they needed to be better in their formative assessment practices, because they were used to traditional forms of assessment, not formative assessment. Based on this, the researcher concluded that as the participants were not educated through formative assessment methods, this change for the teacher to adopt a more formative perspective will take time.

In addition to this, Hatipoğlu (2015) studied with 124 pre-service teachers at Middle East Technical University in Turkey. The aims of the study were to investigate what pre-service teachers knew about assessment and what their expectations were from their course of English Language Testing and Evaluation. The findings demonstrated that the participant students expected to evaluate, select and write exams and prepare their learners for all types of exams. It was also revealed that the learners had limited assessment knowledge after four years in ELT department. Yüce (2015) also studied with 133 pre-service English language teachers from two universities in Konya, Turkey. She

examined pre-service English teachers' conceptions of assessment. The participants were asked which assessment practices they wanted to utilize when they graduated from the university. The data were gathered through the Short version of Teacher Conceptions of Assessment Scale (TCoA-IIIA) that was developed by Brown (2008). 27 items in the scale were classified under improvement, school accountability, irrelavance, and a checklist. The findings revealed that the participants regarded assessment as a means of improving the quality of education. Though they thought that assessment is a means of improving the quality, very surprisingly, for most of them, using assessment was seen as irrelevant. What is more, they favored alternative means of assessment rather than traditional forms of assessment, and they wanted to make use of alternative assessment more when they became teachers.

Another study belongs to Öz and Atay (2017) who investigated the Turkish EFL teachers' perceptions towards in-class language assessment and its link with their classroom practices. The participants were 12 teachers, eight females and four males. The data were obtained through semi-structured interviews. The findings revealed that the teachers were familiar with the basic terms related to classroom assessment; however, when it comes to the practice, they had difficulty in reflecting their assessment knowledge into their classroom parctice. Hence, it was concluded that there was an imbalance between the teachers' assessment literacy and their classroom practices.

Finally, Mede and Atay (2017) made use of the online language testing and assessment questionnaire adapted from Vogt and Tsagari (2014) in order to collect data. The aim was to find out the training needs and practices of Turkish EFL teachers. Both quantitave and qualitative data were utilized. There were 350 teachers (153 males and 197 females) participating in this study from four state and seven private universities in Turkey. The findings showed that the teachers had limited assessment literacy, and they needed training in many areas of testing and assessment, especially the terms related to assessment and classroom-based assessment. They also stated that they were not competent with testing productive and receptive skills. The only areas they felt competent with were grammar and vocabulary. The reason could be that in Turkey, the teachers are fairly familiar with teaching grammar and vocabulary, and accordingly testing them.

To sum up, the studies that were carried out related to assessment literacy levels of teachers aimed to identify their levels in the light of the Standards. These kinds of studies are more in the field of education and psychology. Though less, assessment literacy levels

of EFL teachers have been investigated as well. Rooted in the term assessment literacy, language assesment literacy levels of English teachers have been investigated recently. As there is no instrument to measure it yet, the studies are mostly concerned with the needs of English language teachers with regard to language assessment, proving the inefficiency of pre-service education and lack of professional development, and with the self-reports of the participants related to their assessment knowledge or practices. The number of studies investigating assessment literacy and more specifically language assessment literacy of teachers decreases in Turkish context, unfortunately. Mostly, the studies that were carried out with English teachers mainly investigated the perceptions of English teachers regarding language assessment.

3. METHODOLOGY

3.1. Research Design

The current study was based on a mixed-method design with both quantitative and qualitative data collection elements, putting the former at the center of the data collection and analysis process. Creswell (2012) argued that "the uses of both quantitative and qualitative methods, in combination, provide a better understanding of the research problem and question than either method by itself" (p. 535). For this reason, it is believed that benefitting from two different data collection methods and combining them in this current study provided a better understanding regarding various aspects of the language assessment knowledge of the participants.

According the Dörnyei (2007), mixed-method research, having both quantitative and qualitative components, might potentially result in nine combinations of these components based on their sequence and dominance throughout the data collection and analysis. Among these combinations, this study followed the QUAN ➔ qual combination that refers to the sequential design (➔) of both elements, quantitative having more dominance (QUAN). Dörnyei (2007) maintained that this sequential use of both data collection methods provides both "micro and macro perspectives" (p. 173) regarding the phenomena under investigation; quantitative research for the large-scale tendency, and qualitative research for the micro-level analysis of the research matter by individuals. Based on this mixed-method combination, this study aimed to provide both a general picture of the language assessment knowledge of the EFL teachers in higher education setting in Turkey based on the QUAN part and a micro-level understanding of language assessment by individuals based on the qualitative part. Based on this research design, the quantitative data were collected and analyzed first through the language assessment knowledge scale (LAKS), which was developed and used as the main data collection tool of this study, to reach a general picture regarding the language assessment knowledge of the participants. After that, qualitative data were collected through open-ended questions to get in-depth data and present extended findings on the phenomena under investigation. Finally, all the findings derived from both quantitative and qualitative data were interpreted in the light of the literature and the contextual factors. The following figure illustrates how the data was collected based on the mixed-method design.

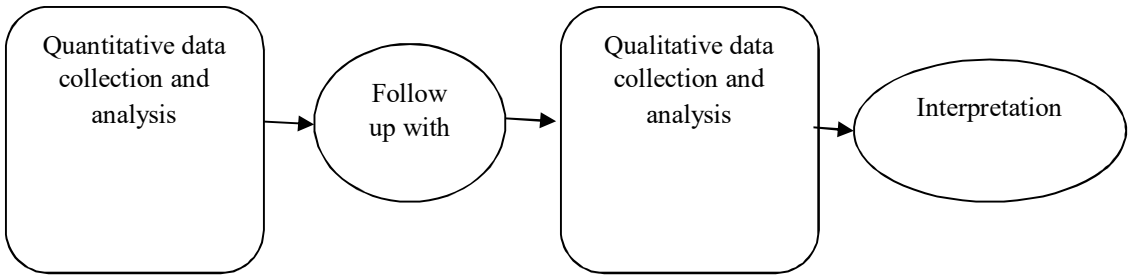

Figure 3.1. *Data collection process (Creswell, 2012, p. 541)*

3.2. Research Context

Turkey is an EFL context in which English does not have an official status. It is taught as a foreign language at primary, secondary and university levels. With the English preparatory programmes they have, schools of foreign languages at universities are the institutions in which English is taught in a systematic and intensive way in Turkey (Aydın, et. al., 2017). The students in these programmes are comprised of different learner profiles. There are three kinds of learners who get education in these programmes. First group includes learners who are going to be educated through English as the medium of instruction in their departments. All students in the programme have to take the proficiency exam at the beginning of the academic year, and if they get the criterion score, mostly 60, or over, they have the right to start their education in their departments. However, if they cannot get the criterion score, they have to expose to an intensive English programme throughout a year, and at the end of the academic year they take the proficiency exam again. If they get the criterion score or over, they can go on their education in their departments for the next academic year because they are considered as proficient by the preparatory programmes. However, if they cannot get the criterion score, they have to repeat the preparatory programme for the next academic year. Second group of learners include the ones whose medium of instruction is 30% English in their departments. They have the same criteria in their preparatory programmes like the students of English medium instruction. Third group of learners are optional preparatory programmes. As this programme is optional, each learner has the right to be a student in preparatory programmes if s/he is willing.

Though there are different learner profiles, what is not changed in these programmes is the existence of testing and assessment. As part of their programmes, all types of learners in preparatory programmes are assessed at regular intervals via quizzes

and exams. Although assessment has an important place in their programmes and language teachers follow some ongoing assessment procedures through the academic year like portfolios, learners are usually and formally tested by quizzes or/and exams. These teachers are responsible for the testing and assessment of these learners, and they are expected to prepare these quizzes or exams. In most of the programmes, there are separate offices such as curriculum office, material office, etc. Testing office is one of them, and language teachers can volunteer to be members of the testing office or the language teachers are assigned responsibilities by the director of the programme (Aydın, et. al., 2017). In most of the programmes, testing offices may include language teachers who have no or little experience in testing. As a member of testing office, language teachers are expected to construct items for separate language skills, conduct exams, evaluate the answers of learners and give a score for the answers. To conduct all such duties, they do not have to take part in teacher training seminars in testing and assessment or they do not have to be knowledgeable in testing and assessment in most of the programmes. They could participate in teacher training programmes or conferences related to assessment, but participation is not obligatory or participating in them is not a prerequisite to be a member of testing office.

Though general regulations such as the length and acceptance to the programme are determined by the council of Higher Education, the implementation of these programmes regarding the curriculum or testing and assessment are determined by the schools themselves. All courses, curricula activities and testing and assessment practices are conducted by the teachers with the guidance of coordinators and head of the school.

3.3. Participants

The current study aimed to present a general picture of the language assessment knowledge of EFL teachers working at universities, and considering these language teachers with various backgrounds, the study aimed to reach to the population without using any sampling strategy. The population of this study included Turkish EFL teachers working at schools of foreign languages at universities in Turkey. The online version of the scale was sent to all the language teachers of the universities which have English preparatory programmes, and they were asked to fill in the scale. The ones who responded to the scale were included as the participants in the study.

As for the language teacher profiles in these programmes, language teachers have diversity in their educational background. In Turkey, a graduate of English Language Teaching, English Language and Literature, English Linguistics, Translation and Interpreting (English) or any departments related with these areas can be a language teacher in preparatory programmes of the universities. Along with the department they graduate from, there are also two criteria to be met; the required scores gotten from ALES and YDS. ALES is an exam for academicians, and is comprised of questions in Turkish language and Maths. YDS is an exam which determines the proficiency level in English. All of the graduates from these departments with satisfactory scores from ALES and YDS have the opportunity to be language teachers in preparatory programmes of the universities.

However, only the department of English Language Teaching is specifically designed to educate pre-service language teachers. As the focus is on educating language teachers and preparing them for their future careers in English language teaching, the courses are designed to serve their purposes. These courses can be grouped as the ones which increase language proficiency such as reading and writing; the ones which give pedagogical knowledge such as methodology and testing courses, and the ones which include linguistic knowledge and literature. In other departments, the students are not educated to be teachers. The students in the other departments have to take courses on pedagogical content knowledge for a short period of time after they graduate. Based on the courses they take, these graduates can begin their careers. As the time is short and the course contents cannot be covered in detail in such a period of time, they have limited exposure to how to teach English to learners, how to teach separate language skills to learners, how to manage classroom, and how to assess learners, etc.

Among 122 universities (85 state and 37 private universities) in Turkey, the scale was sent to the ones with English preparatory programmes. Among these universities, which were decided as the context of the study, 37 state and 16 private universities contributed to the data collection process of the study. The distribution of the universities and the number of the participants are presented in the following table.

Table 3.1. The number of the participants according to universities and regions

REGIONS	UNIVERSITIES	NUMBER OF TEACHERS
Black Sea Region	University 1 (state)	13
	University 2 (state)	2
	University 3 (state)	2
	University 4 (state)	23
	University 5 (state)	3
	University 6 (state)	8
	University 7 (state)	4
Aegean Region	University 8 (state)	2
	University 9 (state)	2
	University 10 (state)	3
	University 11 (state)	6
	University 12 (private)	26
	University 13 (private)	13
	University 14 (state)	8
	University 15 (state)	5
Central Anatolia Region	University 16 (state)	115
	University 17 (state)	4
	University 18 (state)	16
	University 19 (state)	5
	University 20 (state)	2
	University 21 (state)	33
	University 22 (state)	17
	University 23 (state)	11
	University 24 (private)	9
	University 25 (state)	10
	University 26 (state)	9
	University 27 (private)	19
	University 28 (private)	5
	University 29 (private)	17
Mediterranean Region	University 30 (state)	3
	University 31 (state)	2
	University 32 (state)	7
	University 33 (state)	8
	University 34 (state)	2
Marmara Region	University 35 (state)	9
	University 36 (state)	4
	University 37 (state)	9
	University 38 (state)	3
	University 39 (private)	13
	University 40 (private)	2
	University 41 (private)	2
	University 42 (private)	9
	University 43 (private)	2
	University 44 (private)	3
	University 45 (private)	2
	University 46 (private)	9
	University 47 (private)	30

Southeastern Anatolia Region	University 48 (state)	2
	University 49 (state)	2
	University 50 (state)	5
	University 51 (state)	4
	University 52 (private)	10
Eastern Anatolia Region	University 53 (state)	8
TOTAL	**53**	**542**

In total, the participants included 542 teachers from 53 universities. In addition to the institutional distribution, the participant teachers had also diversity in terms of their demographic features stated in the scale. These demographic features and the number of the teachers having these features are shown in the following table.

Table 3.2. *Demographic features and the number of the participants*

Demographic feature	Number of the Participants	Percentage
Gender	Male – 174	32
	Female - 368	68
Years of experience	1-5 years – 86	16
	6-10 years – 173	32
	11-15 years – 114	21
	16-20 years – 100	18
	More than 21 – 69	13
Educational background	BA – 238	44
	MA – 255	47
	PhD – 49	9
The BA programme graduated	ELT – 347	64
	Non-ELT - 195	36
The current workplace	State University – 372	68
	Private University – 170	32
Being a testing office member	Yes – 260	48
	No – 282	52
Had a separate testing/assessment course in pre-service	Yes – 260	48
	No – 282	52
Attended any trainings on language testing/assessment	Yes – 282	52
	No – 260	48

The last step of the data collection process included the qualitative phase. For this, seven open-ended questions (See Appendix B), which were focusing on the findings from the quantitative data and teachers' needs in language assessment, were sent via e-mails to 20 teachers, 10 testing members and 10 non-testers. 10 teachers worked at state universities whereas the other 10 worked at private universities. These teachers were purposefully determined from different universities so that they could provide more and

richer data for the foci of the questions to get a micro-level understanding of the research focus. They were asked to answer the questions in detail giving personal and context-specific explanations. Among them, 11 teachers responded to the email and answered all the questions completely. Six of these teachers were the members of the testing office. Out of six teachers, three of them were working at state universities, on the other hand, three of them were working at private universities. Five of these teachers were not the members of the testing office, two of whom were working at state universities; however, three of them were working at private universities.

3.4. Data Collection and Analysis Process

3.4.1. Developing language assessment knowledge scale (LAKS)

Due to the paucity of research examining the language assessment knowledge of EFL teachers and lack of a valid data collection instrument to conduct such studies, a valid and reliable data collection instrument was developed within the scope of the study. The process towards the development of the instrument is explained in detail as follows. First of all, in order to provide a deep theoretical background to the instrument, the books referenced so far on language testing and assessment (Harris, 1969; Hughes, 1989; Heaton, 1990; Bachman, 1990; Bailey, 1998; Alderson, 2000; Buck, 2001; Weigle, 2002; Brown, 2003; Luoma, 2004; Fulcher & Davidson, 2007, 2012; Coombe, et. al., 2012, etc.) were read by the researcher. While reading, all the knowledge elements, which were stated in those books as "need-to-know" about testing or assessing language skills; that is, reading, listening, writing and speaking, were listed by the researcher. Then, the researcher chose the ones repeated in references mentioned above for the item pool for each language skill. This list consisted of 237 items in total (49 items for reading, 61 for listening, 74 for writing, and 53 for speaking). Next, three experts with a PhD degree in ELT went over the item pool in detail focusing carefully on the comprehensibility and orthography of the items and the compatibility of each item for the language skill it was listed in. At the end of this initial step, 17 items (3 items from reading, 2 from listening, 6 from writing, and 6 from speaking) were removed from the instrument, and the very initial format of the scale had four constructs; assessing reading (46 items), assessing listening (59 items), assessing writing (68 items) and assessing speaking (47 items), consisting 220 items in total.

At the second stage, individual meetings with 10 teachers having various years of teaching experience and educational background from the school of foreign languages of different universities were held. In these individual meetings, the teachers were asked to read the items and make comments on whether the items were clear to them and they had any difficulty in understanding the terminology in the items. At the end of those meetings, no item was removed from the list but several revisions were made based on the suggestions provided by the teachers to make the wording clearer for further stages.

The third stage of developing the instrument included the expert opinion process. For this, the instrument was designed in a questionnaire format having four different parts, each for a different language skill and the items were listed in these parts. For each item, the researcher put three choices as "necessary, not necessary, needs revision (please justify)" similar to a Likert-scale, and the items were provided to the experts, 14 academicians who studied on testing and assessment or gave related courses in higher education level in the fields of English language teaching and testing and evaluation at different universities. In one month, 11 of the experts responded to the initial format of the instrument, and provided feedback on each item. Based on the suggestions provided by these experts, 67 items were removed from the instrument, and some revisions were made on several items. At the end of all these stages, 153 items remained in the scale (reading: 37 items; listening: 33 items; writing: 48 items, and speaking: 35 items).

At the fourth stage of the process, the scale was presented to real practitioners, which was believed to contribute significantly to the validation of the instrument, and a meeting was organized with the testing office members of an English preparatory programme of one of the leading universities in Turkey. The meeting included 18 teachers, 6 of them had PhD or MA in testing and evaluation or in ELT. They were sent the instrument before the meeting, and were asked to respond to and comment on it beforehand. The meeting in which the participants and the researcher discussed the validity, comprehensibility and compatibility of each and every item lasted about five hours. At the end of the meeting, which provided the researcher a deeper insight from the perspectives of the practitioners, 41 items were removed from the instrument, and several revisions were made on the remaining ones. Finally, the instrument which is called Language Assessment Knowledge Scale (LAKS, henceforth) consisting of 112 items (reading: 28 items; listening: 26 items; writing: 34 items, and speaking: 24 items) were ready for the piloting process.

At the fifth stage, 112 item were piloted with 50 teachers who were then excluded from the actual study. They were asked both to complete the questionnaire and make comments on it. However, after receiving their answers, it was seen that there occurred some problems with the statistical analyses of the scale, and no model came out as a result of the analyses. When the reasons of these were investigated, it was observed that the participants tended to give the same answers (all true or all false) towards the end of the scale, and some of the participants did not even finish completing. Furthermore, the comments made by those participants revealed that there were too many items to respond in the scale and it took too much time, demotivating them to complete. Many participants sent e-mails to the researcher stating that they could not concentrate on the items because there were too many items in the scale. Based on those feedbacks, which consisted of the elements that had the potential to influence the validity and the reliability of the scale negatively, five academicians who were experts in ELT analysed the scale in detail and made a comment for each and every item in each skill on whether this item should be in the scale or not. Then, they compared and contrasted their comments with each other, and after negotiation, they decided to keep the items that were fundamental for a language teacher to know regarding the language assessment of a foreign language, and the other items were eliminated from the study. While removing the items, there were also other criteria that were taken into consideration. One is that the answers of certain items were more predictable than the others, thus, it was thought that they would be less evaluative. "portfolio assessment is product-based" and "combining vocabulary and reading in a single test is avoided" are some of the examples in this group. The second is that some items were longer than the other items, which, in turn increased the cognitive load on the readers; so, some of these items were removed from the scale such as "using options that involve opposing ideas in the same multiple choice question does not pose a problem". The next one was the items in which the terms used could not be understood very easily. In other words, there were some terms which might lead the readers to confusion. An example for this is "using controlled vocabulary in listening texts has advantages over free vocabulary", for which, the participants had difficulty understanding what was meant by controlled or free vocabulary. The other one was the dominance or repetition of some topics in a specific scale. For instance, in assessing writing, there were too many items related to holistic scale and analytic scale; hence, some of them were removed from the scale. As a result, in line with the comments and the answers of the participants in the

piloting, and after the experts' negotiation, 52 items were removed from the scale, and the remaining 60 items, with 15 in each construct were sent to all the language teachers working at the schools of foreign languages in Turkey in an online platform. This removal and revision process in all these stages is shown in the table below.

Table 3.3. *Revision process of the scale*

	reading	**listening**	**writing**	**speaking**	**in total**
1st stage	49	61	74	53	237
(Three experts with PhD in ELT checking for comprehensibility)	-3	-2	-6	-6	-17
2nd stage	46	59	68	47	220
(Checking with 10 teachers)	-	-	-	-	-
3rd stage	46	59	68	47	220
(Expert opinion)	-9	-26	-20	-12	-67
4th stage	37	33	48	35	153
(Training with the testing office members)	-9	-7	-14	-11	-41
5th stage	28	26	34	24	112
(Piloting with 50 teachers and expert opinion)	-13	-11	-19	-9	-52
Final Version	15	15	15	15	60

3.4.2. Data collection of the main study

The data of this study were collected during the early days of the spring semester of 2017-2018 academic year. After the development and validation process of LAKS at the end of the first term, it was sent in an online format to all the teachers working at the school of foreign languages at universities throughout the country. During data collection process, the researcher sent reminder e-mails to the participants and the head of their schools in order to encourage the participants to respond to the scale. The process in which the quantitative data of the study was collected lasted about one and a half month and at the end of this period 542 participants responded LAKS completely, and these participants formed the core data of this study.

In addition to this, the researcher prepared open-ended questions based on the findings from the quantitative data and asking teachers' needs in language assessment to provide more in-depth data regarding the language assessment knowledge of teachers participating this study. The questions were checked by three experts in the field of ELT to make them more valid and to-the-point. Moreover, the researcher also asked three language teachers to check the orthography of the items. These open-ended questions

were sent to 20 language teachers from different universities via e-mail and 11 of them responded to the e-mail answering all the questions completely.

3.5. Data Analysis

The first research question of the study focused on the psychometric properties of LAKS. After the data were collected from 542 participants, first the statistical analyses revealing the psychometric properties of LAKS were conducted. So as to conduct the statistical analyses, confirmatory factor analysis, second order confirmatory factor analysis, item correlation, and Cronbach alpha were utilized. This analysis process including the types of analyses is presented in Table 3.4 in detail.

The aim of this scale-development was to measure the knowledge of EFL teachers in language assessment in general, and in assessment of four language skills (reading, listening, writing, speaking) in particular. For this purpose, 15 items were determined for each sub-construct (each skill) after an elaborated and comprehensive validation process, as explained in the previous part. The items were designed in “true, false, don’t know” format, and the participants were rated "1" if their answers to these items were correct according to what assessment literature suggests, and "0" if they were incorrect or they chose “don’t know”. Based on this, the highest score that can be achieved for each sub-construct is 15, and the score for the total is 60 in the scale.

The two main features that are expected to take place in a measurement tool are reliability and validity (Dörnyei, 2007). In general, it is possible to determine the reliability of a measuring tool, which is defined as the degree of accuracy (Gelbal, 2013), through several ways. For this study, the Cronbach Alpha coefficient was reported for each construct to ensure internal consistency. Validity, on the other hand, is defined as the serving level of the measuring instrument for the purpose (Gelbal, 2013). It can be said that demonstrating the validity of the scale is a process rather than a single analysis. The content validity of the LAKS was ensured through several stages; review of the literature, opinions of various groups of experts in the writing of the items and the piloting process can be used to demonstrate that the contents of the items serve their purpose.

Factor analysis, which is usually applied to demonstrate the construct validity, is also used for the empirical findings on the validity of the scales. In scale studies, factor analysis is divided into two; Exploratory Factor Analysis (EFA) which is used to determine the constructs when there is no “a-priori knowledge” about the factor structure

of the scale, and Confirmatory Factor Analysis (CFA) which is applied to validate the factor structure as a model when there is strong prior knowledge about the construct of the scale (Çokluk, Şekercioğlu, & Büyüköztürk, 2012). Within the scope of this study, since four theoretical sub-constructs, (assessing reading, assessing listening, assessing writing, assessing speaking) and competencies related with those skills were clearly identified as sub-dimensions of language assessment in the literature, they were regarded as the constructs of the scale. In addition to this, each item for these sub-constructs was written based on the resources and seminal works published on skill-based assessment. For this reason, it can be said that the scale had a very strong a-priori. Therefore, the second order Confirmatory Factor Analysis (CFA) was applied to establish the validity of the Language Assessment Knowledge (LAK) scale.

On the other hand, the main quantitative data collected during the early days of the spring semester of 2017-2018 on teachers' language assessment knowledge were analyzed through descriptive and inferential statistics. The following table presents the statistical methods that were used to analyze the quantitative data of this study in line with the research questions.

Table 3.4. *Statistical methods used in analysis*

The focus of the research question	**The statistical method**
R. Q. 1. Psychometric properties of LAKS	Confirmatory factor analysis Second order confirmatory factor analysis Item correlation Cronbach alpha
R. Q. 2. The level of general and skill-based language assessment knowledge	Descriptive statistics (mean, percentage, standard deviation, etc.) One sample t-test
R. Q. 3. The relationship of their skill-based knowledge	Pearson Correlation
R. Q. 4. The impact of demographic features on the knowledge level of participants	Inferential statistics (Independent samples T-test, one-way ANOVA)
R. Q. 5. Perceived self-competency and LAK level	One-way ANOVA

First of all, the level of language assessment knowledge of the participants was presented through descriptive analyses in terms of their overall and skill-based scores. In addition to this, the impact of the demographic features of the participants on the level of their language assessment knowledge was also analyzed via inferential statistics. The

impact of the demographic features consisting of two independent groups such as workplace or having a testing course in BA were determined through independent-samples T-test whereas one-way ANOVA was used to determine whether there is a significant difference between three or more independent groups (educational background, years of experience, etc.). Finally, the relationship between the skill-based knowledge was investigated to find out whether there is a positive or negative correlation among the skills and the overall knowledge through Pearson correlation.

On the other hand, the qualitative data derived from the open-ended questions were analyzed based on the qualitative content analysis scheme of Creswell (2012). The answers of the participants were broken into chunks and code-labelled by the researcher. Finally, certain themes based on these initial codes were identified. At the end, the emerging themes were presented in frequencies. The following figure presents a systematic representation of the qualitative data analysis process.

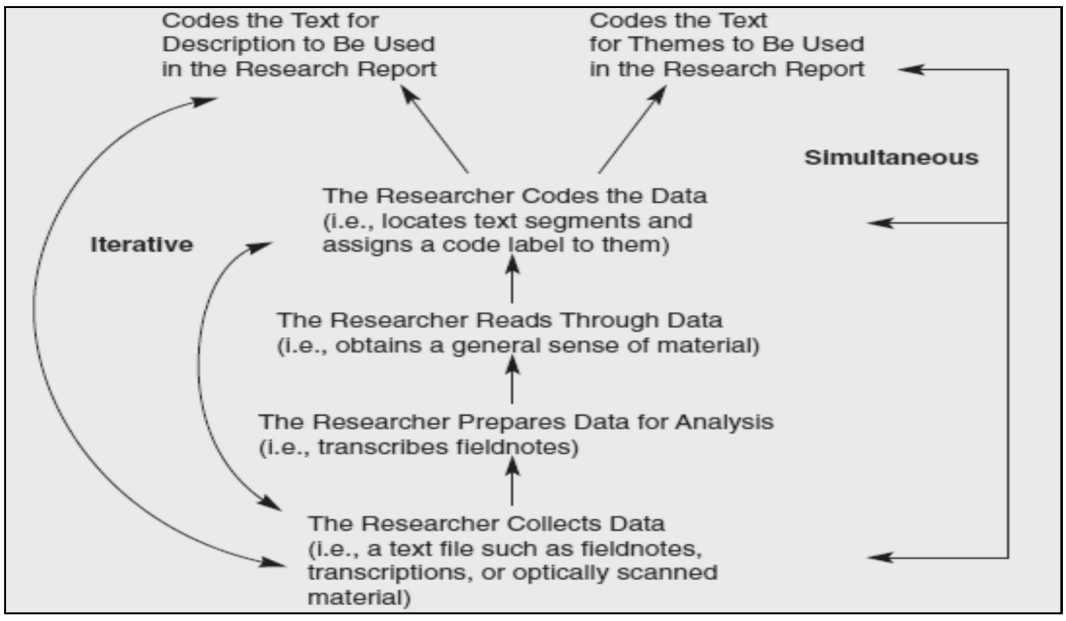

Figure 3.2. *Qualitative data analysis scheme (Creswell, 2012, p.237)*

To increase the validity of qualitative studies, there are some steps that should be taken into consideration in the literature. Triangulation, which is basically defined as "the use of multiple, independent methods of obtaining data in a single investigation in order to arrive at the same research findings" (Mackey & Gass, 2005, p. 181) is one of the most important elements in qualitative research. Agreeing with the definition of triangulation

above, Yıldırım and Şimşek (2016) also defined triangulation as including many participants in a study who have various features and background for gathering richer data. During the qualitative data collection and analysis process, the researcher followed two steps to ensure triangulation. First, for data collection, the researcher chose from different participants from different contexts such as selecting participants from both state and private universities, and participants who were testing members or not. Besides, both quantitative and qualitative data collection instruments were utilized to obtain data from the participants. Additionally, a colleague holding a PhD in ELT assisted the data analysis process while coding and identifying the emerging themes in order to increase the interrater reliability of the data analysis. Both raters analyzed the answers of the participants to open-ended questions independently, and they came up with some codes, and eventually certain themes. Then, they compared and contrasted their analysis with each other, and they had 80% agreement on labelling these codes and themes. The labelling of the remaining 20% were agreed through negotiation. All those steps contributed to the triangulation of qualitative data collection and analysis process.

4. FINDINGS

4.1. Psychometric Properties of LAKS

The first research question aimed to reveal the psychometric properties of LAKS. First, in order to confirm the compatibility of the items with the constructs (assessing reading, assessing listening, assessing writing and assessing speaking), and the compatibility of these constructs with language assessment knowledge; in other words, the model data fit in general, second order CFA was performed using the Mplus 7.0 package programme. Since the responses given for each item were categorical, WLSMV was used as the proficiency estimator. Since CFA is included in the structural equation modeling family, the model data compatibility was first investigated for the results of CFA. The results and interpretations are as follows.

Table 4.1. *Model-fit indices derived from second order CFA*

Fit Indice	Reference points	Value	Comment
Chi-square/df	2.5 – 5 good fit < 2.5 perfect fit	1.41	Perfect Fit
RMSEA	.08 - .05 good fit < .05 perfect fit	.028	Perfect Fit
CFI	.90 - .95 good fit > .95 perfect fit	.981	Perfect Fit
TLI	.90 - .95 good fit > .95 perfect fit	.980	Perfect Fit

In the structural equation modeling studies, the expected chi-square value is not significant, in other words, the value of "p" must be bigger than .05. However, this value can be misleading because it is sensitive to the size of the sample. For this reason, the value obtained by dividing the chi-square by degrees of freedom is generally reported. At this point, the value of the model which is below 2.5 indicates a perfect fit. In the second order CFA, the RMSEA value between .08 and .05 indicates a good fit, and the values smaller than .05 indicate a perfect one. Moreover, a good fit for CFI and TLI values is between .90 and .95, and a perfect fit is for values above .95 (Hu & Bentler, 1999; Byrne, 2012; Çokluk, Şekercioğlu, & Büyüköztürk, 2012). At this point, .028 as the RMSEA value, .981 as the CFI and .980 as the TLI value of the scale revealed a perfect fit in this

study. Thus, it can be said that the complete statistics obtained are indicative of a perfect model data fit.

Standardized values in the structural equation modeling are interpreted as standardized coefficients in the regression. In the context of CFA, these values are seen as factor loadings. Factor loadings for each item, standard errors and t values for these values are presented in the table below.

Table 4.2. *Factor loadings for each item*

Factor	Item no	Factor loading	SE	t	R-square
Assessing Reading	1	0.869*	0.021	41.055	0.755
	2	0.536*	0.040	13.342	0.287
	3	0.872*	0.019	46.981	0.760
	4	0.777*	0.022	35.885	0.603
	5	0.064	0.058	1.110	0.004
	6	0.753*	0.031	24.012	0.567
	7	0.842*	0.020	41.343	0.708
	8	0.895*	0.022	40.634	0.801
	9	0.632*	0.039	16.402	0.400
	10	0.279*	0.053	5.238	0.078
	11	0.869*	0.019	46.619	0.756
	12	0.805*	0.020	39.951	0.648
	13	0.990*	0.018	55.177	0.980
	14	0.485*	0.047	10.282	0.235
	15	0.855*	0.027	31.767	0.730
	16	0.470*	0.045	10.354	0.221
	17	0.841*	0.021	40.876	0.707
	18	0.824*	0.021	39.164	0.679
	19	0.266*	0.058	4.587	0.071
	20	0.730*	0.024	30.613	0.533
	21	0.680*	0.039	17.548	0.463
Assessing Listening	22	0.631*	0.027	23.514	0.398
	23	0.697*	0.024	29.423	0.486
	24	-0.043	0.059	-0.728	0.002
	25	0.021	0.059	0.351	0.000
	26	0.902*	0.018	48.991	0.814
	27	0.576*	0.033	17.316	0.332
	28	0.262*	0.052	5.025	0.069
	29	0.660*	0.026	25.784	0.435
	30	0.969*	0.020	48.969	0.940
	31	0.121	0.066	1.847	0.015
	32	0.889*	0.031	28.386	0.791
	33	0.078	0.062	1.267	0.006
	34	-0.046	0.071	-0.644	0.002
Assessing Writing	35	0.431*	0.054	7.975	0.186
	36	0.013	0.062	0.207	0.000
	37	0.025	0.076	0.330	0.001
	38	-0.018	0.066	-0.269	0.000
	39	0.631*	0.045	14.132	0.398
	40	0.585*	0.044	13.245	0.343
	41	-0.033	0.062	-0.530	0.001

	42	0.442*	0.051	8.680	0.195
	43	0.085	0.062	1.371	0.007
	44	0.613*	0.046	13.263	0.376
	45	0.609*	0.042	14.467	0.371
	46	0.077	0.059	1.301	0.006
	47	-0.019	0.067	-0.290	0.000
	48	0.076	0.061	1.257	0.006
	49	0.479*	0.044	10.964	0.229
	50	-0.027	0.059	-0.463	0.001
	51	0.255*	0.061	4.194	0.065
	52	0.684*	0.033	20.660	0.468
Assessing Speaking	**53**	0.304*	0.052	5.803	0.092
	54	0.110	0.063	1.759	0.012
	55	0.916*	0.018	50.230	0.839
	56	0.350*	0.053	6.616	0.123
	57	1.020*	0.014	73.086	1.00
	58	0.845*	0.026	32.142	0.715
	59	0.747*	0.033	22.386	0.557
	60	0.039	0.063	0.613	0.002

* $p < .05$

The values given in the first column above are referred as standardized path coefficients, and these values are accepted as factor loadings in CFA. The coefficients are valued between -1 and +1, and the higher the value is, the higher its relationship with the latent variable is. The second column refers to the standard error values and the third column includes the t values, which are obtained by dividing the factor loading of an item to its standard error. Getting higher t values increases the significance of the items. The last column gives the R-square values which equal to the square of factor loadings. This value is between 0 and 1, and as it gets closer to 1, the amount of variance explained in the observed variable increases. Based on these explanations, it can be seen that the factor loadings of most of the items in assessing reading and assessing listening are significant and satisfactory whereas there exist several items with low factor loadings in assessing writing and assessing speaking.

In the next step, the structural values obtained were reported on the model. This figure is shown below.

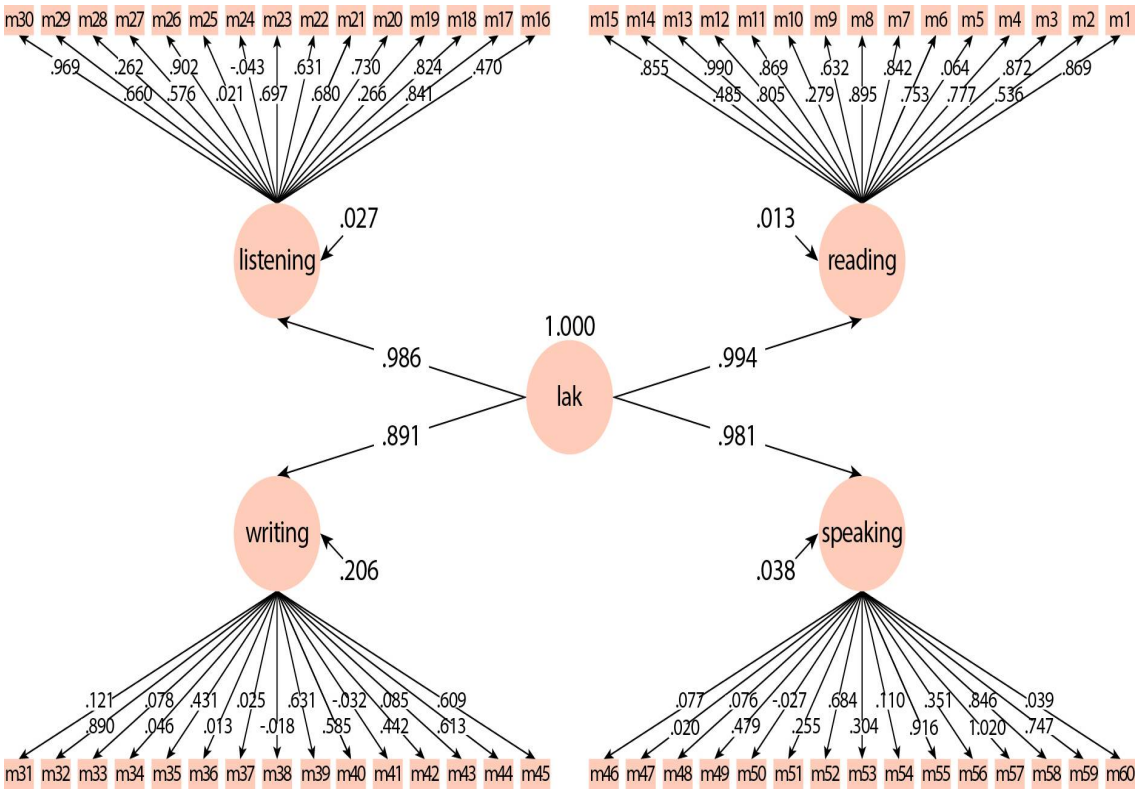

Figure 4.1. *Results of the second order CFA*

Based on the second order CFA, the figure above reveals how LAK explains its constructs (assessing reading, assessing listening, assessing writing, and assessing speaking) in terms of their variance. Firstly, standardized path coefficients were reported as .98 for assessing listening, .99 for assessing reading, .89 for assessing writing and .98 for assessing speaking. That means, one standard deviation change in LAK (1.000) would lead to .986 standard deviation change in assessing listening, 0.994 standard deviation change in assessing reading, .891 standard deviation change in assessing writing, and .981 standard deviation change in assessing speaking, all of which are good indicators of variance explanation. In addition to this, the error variance values were found as .027 for assessing listening, .013 for asseesing reading, .206 for assessing writing and .038 for assessing speaking. In other words, these values mean that LAK explains 97% variance of assessing listening, 98% variance of assessing reading, 80% variance of assessing writing and almost 96% variance of assessing speaking. In short, as all these values suggest, the model presents a perfect model data fit in terms of explaining LAK and its constructs.

4.1.1. Reliability analysis

In developing and validating measurement instruments, presenting the statistical values related with the reliability is another important factor. The following table gives the Cronbach Alpha coefficients of LAKS in total and its sub-constructs.

Table 4.3. *Reliability analysis for Language Assessment Knowledge Scale (LAKS) and its sub-constructs*

Constructs	Cronbach Alpha
Language Assessment Knowledge Scale	.91
Assessing Reading	.88
Assessing Listening	.78
Assessing Writing	.49
Assessing Speaking	.65

The findings above reveal that Cronbach Alpha coefficient of LAKS in total was .91 which is a highly satisfactory value, and it shows that LAKS has a statistically high reliability to be used as a measurement tool. When the table is examined, it is also seen that the Cronbach Alpha value for assessing reading appeared to be .88, again referring to a high level of reliability. The Cronbach Alpha coefficient for assessing listening sub-construct was obtained as .78, which means that the scale has internal consistency at an acceptable level. The Cronbach Alpha coefficient calculated for assessing writing sub-construct was found out as .49. Since the value is below the critical limit of .60, this construct resulted in a lower confidence in internal consistency. Finally, the Cronbach Alpha coefficient for assessing speaking was found to be .65, which is again an acceptable value for the internal consistency. In addition to the reliability values above, item-total correlations related with each item were also calculated under each skill. The following table presents the values for the items.

Table 4.4. *Item-total correlation coefficients of the items under each skill*

Item No	Item-Total Correlation			
	Assessing Reading	**Assessing Listening**	**Assessing Writing**	**Assessing Speaking**
1	.686	.363	.082	.109
2	.382	.586	.410	-.009
3	.660	.536	.100	.124
4	.564	.204	.009	.282
5	.048	.476	.264	-.037
6	.583	.490	.106	.201
7	.632	.372	.014	.444
8	.720	.449	.106	.176
9	.481	.024	.314	.046
10	.191	.039	.314	.580
11	.699	.625	.040	.284
12	.608	.327	.311	.637
13	.797	.141	.018	.583
14	.312	.439	.235	.484
15	.674	.706	.224	.038

When the item-total correlation values in the above table are examined, it is seen that items 5 and 10 in reading are relatively low in size. The coefficients obtained for other items in the construct of assessing reading were satisfactory. As for the construct of assessing listening, it is seen that the correlation value obtained for items 9, 10 and 13 in this subconstruct was relatively low and the other items had satisfactory values. The third construct was assessing writing and in this construct, it is seen that most of the items had low-level correlation values. Finally, the last construct of the scale was assessing speaking, and it is seen that the item-total correlation values obtained for most of the items are below .50 and relatively low in size.

When all the item-total correlation values are examined, it can be concluded that several items under each construct have a relatively low level of item-total correlation.

However, after this statistical analysis, three academicians who significantly contributed to initial validation process of LAKS were asked to provide expert opinion on those items. Based on these experts' opinions, it was decided that these items were important for the content validity of the scale and their contribution to LAKS in general was significant in terms of measuring language teachers' assessment knowledge. Besides, considering model-data fit and reliability coefficients of the constructs, it can be said that the scale presented satisfactory statistical values with those items. Due to all these reasons, the items with relatively low level of item-correlation values were decided to kept in the scale.

4.2. General and Skill-based Language Assessment Knowledge Level of EFL Teachers

The second research question of the study aimed to investigate general and skill-based LAK level of EFL teachers working at Turkish higher education context. The participants, including 542 teachers from different universities, completed Language Assessment Knowledge Scale (LAKS) and the findings derived from their responses are presented in Table 4.5.

Table 4.5. *General and skill-based LAK level of EFL teachers in Turkish higher education context*

ITEMS	N	True	False	Don't Know	Mean	SD
ASSESSING READING	**(Bold ones refer to the participants with correct answers)**					
1. Asking learners to summarize the reading text is a way of assessing their reading skills.	542	**269**	257	16	,496	,500
2. When asking several questions about a reading text, all the questions are independent of each other.	542	**153**	343	46	,282	,450
3. Cloze test is used for assessing the main idea of the text.	542	230	**250**	62	,461	,498
4. In a reading exam, using a text learners have encountered before is not a problem.	542	278	**190**	74	,350	,477
5. One reading text is enough to be included in a reading exam.	542	108	**400**	34	,738	,440
6. The language of the questions is simpler than the text itself.	542	**264**	220	58	,487	,500
7. Errors of spelling are penalized while scoring.	542	256	**237**	49	,437	,496
8. Taking vocabulary difficulty into consideration is necessary in assessing reading skills.	542	**288**	224	30	,531	,499
9. Including not stated/doesn't say along with true/false items has advantages over true/false items.	542	**236**	221	85	,435	,496

10. The more items a reading text is followed, the more reliable it becomes.	542	**198**	200	144	,365	,481
11. Using the same words in the correct option as in the text is not a problem.	542	241	**243**	58	,448	,497
12. Simplification of reading texts is avoided.	542	243	**205**	94	,378	,485
13. Reading texts in a reading exam include various genres (essay, article, etc.).	542	**328**	188	26	,605	,489
14. In top-down approach, assessment is on overall comprehension of the reading text.	542	**267**	110	165	,492	,500
15. Using ungrammatical distractors in multiple choice questions in a reading exam is a problem.	542	**296**	199	47	,546	,498
READING-TOTAL	**542**				**7,055**	**4,470**
ASSESSING LISTENING						
16. Using reading texts for listening purposes poses a problem.	542	**160**	292	90	,295	,456
17. Including redundancy (e.g. what I mean to say is that ….) in a listening text poses a problem.	542	243	**228**	71	,420	,494
18. Any type of listening text is used for note-taking.	542	267	**223**	52	,411	,492
19. Spelling errors are ignored in scoring the dictation.	542	**92**	400	50	,169	,375
20. Errors of grammar or spelling are penalized while scoring.	542	319	**169**	54	,311	,463
21. A listening cloze test is a way of selective listening.	542	**286**	139	117	,527	,499
22. Phonemic discrimination tasks (e.g. minimal pairs such as sheep-ship) are examples of integrative testing.	542	209	**63**	270	,116	,320
23. Scoring in note-taking is straightforward.	542	253	**132**	157	,243	,429
24. In discrete-point testing, comprehension is at the literal/local level.	542	**199**	45	298	,367	,482
25. Using dictation diagnostically in assessing listening skills does not pose a problem.	542	**172**	171	199	,317	,465
26. Giving learners a transcript of the listening text is a valid way of assessing listening skills.	542	224	**259**	59	,477	,499
27. Dictation is a kind of discrete-point testing.	542	253	**52**	237	,095	,294
28. Inference questions based on intelligence are avoided in listening tests.	542	**100**	399	43	,184	,388
29. Asking learners to listen to names or numbers is called intensive listening.	542	278	**126**	138	,232	,422
30. In selective listening, learners are expected to look for certain information.	542	**315**	187	40	,581	,493
LISTENING-TOTAL	**542**				**4,752**	**3,291**
ASSESSING WRITING						
31. Giving two options to learners and asking them to write about one ensure reliable and valid scoring.	542	312	**160**	70	,295	,456
32. Analytic scoring is used to see the strengths and weaknesses of learners.	542	**279**	177	86	,514	,500

33. The parts of a scoring scale and the scores in each part do not change for different levels of learners.	542	150	**335**	57	,618	,486
34. When there is a disagreement between the scores of the two raters, they score the written work again.	542	381	**134**	27	,247	,431
35. Learners are required to write about at least two tasks in the exam rather than one task.	542	**149**	309	84	,274	,44688
36. Giving restrictive prompts/guidelines to learners for the writing task is avoided.	542	155	**333**	54	,614	,487
37. Giving learners an opinion and asking them to discuss it is a valid way of assessing their writing skills.	542	420	**72**	50	,132	,339
38. Using visuals which guide learners for writing poses a problem.	542	50	**422**	70	,778	,415
39. Holistic scoring is used to see whether the learner is proficient or not at the end of the term.	542	**257**	161	124	,474	,499
40. Analytic scoring leads to greater reliability than holistic scoring in writing.	542	**216**	192	134	,398	,490
41. In controlled writing, learners have the chance to convey new information.	542	163	**261**	118	,481	,500
42. Classroom evaluation of learning in terms of writing is best served through analytic scoring rather than holistic scoring.	542	**214**	167	161	,394	,489
43. Irrelevant ideas are ignored in the assessment of initial stages of a written work in process writing.	542	173	**292**	77	,538	,498
44. Providing a reading text for writing is a way of assessing writing skills.	542	**250**	196	96	,461	,498
45. Mechanical errors (e.g. spelling and punctuation) are dealt with in the assessment of later stages of a written work.	542	**189**	298	55	,348	,477
WRITING-TOTAL	**542**				**6,573**	**2,478**
ASSESSING SPEAKING						
46. When the interlocutor does not understand the learner, giving that feeling or saying it poses a problem.	542	308	**191**	43	,352	,478
47. Giving learners one task is enough to assess speaking skills.	542	34	**486**	22	,896	,304
48. Interlocutors' showing interest by verbal and non-verbal signals poses a problem.	542	125	**386**	31	,712	,453
49. When it becomes apparent that the learner cannot reach the criterion level, the task is ended.	542	**157**	320	65	,289	,454
50. Using holistic and analytic scales at the same time poses a problem.	542	149	**231**	162	,426	,494
51. Reading aloud is a technique used to assess speaking skills.	542	**87**	380	75	,160	,367
52. In interlocutor-learner interviews, the teacher has the chance to adapt the questions being asked.	542	**209**	277	56	,385	,487
53. In interactive tasks, more than two learners pose a problem.	542	**149**	316	77	,274	,446

54. The interlocutor gives the score when the learner is in the exam room.	542	72	**430**	40	,793	,405
55. In a speaking exam, production and comprehension are assessed together.	542	**282**	231	29	,520	,500
56. Asking learners to repeat a word, phrase or a sentence is a way of assessing speaking skills.	542	**112**	359	71	,206	,405
57. Discussion among learners is a way of assessing speaking skills.	542	**312**	213	17	,575	,494
58. A checklist is a means of scoring oral presentations in in-class assessment.	542	**288**	183	71	,531	,499
59. When the focus is to assess discourse, role plays are used.	542	**270**	166	106	,498	,500
60. In peer interaction, random matching is avoided.	542	**100**	342	100	,184	,388
SPEAKING-TOTAL	**542**				**6,808**	**2,784**
LAKS-TOTAL	**542**				**25,190**	**11,390**

The responses of the participants were analyzed through descriptive statistics, and the results showed that the participants' mean score in LAKS was 25 over 60. In other words, the number of the items answered correctly by the teachers were 25,19 on average, which means that their knowledge level in language assessment is lower than 50%. To confirm this, in other words, to reveal whether this mean score is statistically and significantly lower than the half of the total score, one sample t-test was applied. The lowest score that can be obtained from this scale was 0, and the highest score was 60. Thus, the score of 30, which is the half of the total score, was accepted as the reference point, and it was compared with 25,19 (the mean score of the participants). The findings are presented as below.

Table 4.6. *One sample t-test results*

Mean diff.	**df**	**t**	**p**
4.81	541	-9.83	.000*

*p<.05

According to the values above, it was found that the mean difference (4.81) between the participants' mean score (25.19) in the scale and the half of the maximum score (30) is statistically significant. That means their LAK level in general is significantly low. Besides, one sample t-test was also applied for each skill to find out whether the mean score regarding each skill is significantly lower than the half of the total point for each skill. The findings are presented in the table below.

Table 4.7. *One sample t-test results – skill based*

	Mean diff.	df	t	p
Assessing Reading	-,44	541	-2,31	.021*
Assessing Listening	-2,74	541	-19,42	.000*
Assessing Writing	-,926	541	-8,69	.000*
Assessing Speaking	-,691	541	-5,78	.000*

*p< .05

There were 15 items in each skill. The minimum and maximum scores for each skill were 0 and 15. Thus, the half of the total point was 7,5. The mean scores for each skill were 7,055 for assessing reading, 4,752 for assessing listening, 6,573 for assessing writing, and 6,808 for assessing speaking. The results shown in the table above revealed that the participants' mean scores in each skill were significantly lower than the half of the total score.

In addition to the mean score of LAKS in total providing a general picture regarding the LAK level of the participants, their knowledge level based on each language skill was also examined. The findings displayed that though their mean score was again less than the half of the total number of questions in reading section, the participants had the highest mean score in assessing reading (7,055 over 15) which means that they know more about assessing reading compared to assessing other skills. Moreover, among the items measuring the knowledge of the participants in assessing reading, "one reading text is enough to be included in a reading exam (False)" received the highest mean score whereas the item "when asking several questions about a reading text, all the questions are independent of each other (True)" was the least correctly answered one by the participants. Finally, based on the mean score of each item in this section, it can be said that the participants might have the highest mean score in assessing reading, but it is clear that they still have certain weaknesses in terms of their knowledge in assessing this skill. In LAKS, the items between 16 and 30 aimed to measure the knowledge of the participants in assessing learners' listening skills. According to the results in the table, the participants got a mean score of 4,752 over 15 and listening was found to be the skill in which the participant teachers were the least knowledgeable in terms of language

assessment. In terms of the items in this section, “in selective listening, learners are expected to look for certain information (True)” was answered correctly by more than half of the participants and received the highest mean score in assessing listening part whereas “dictation is a kind of discrete-point testing (False)” had the lowest mean score and was answered incorrectly or not known by more than 90% of the teachers. Considering the mean scores in this section in general, it can be concluded that the participants had weaknesses regarding each and every item in assessing listening.

The items between 31 and 45 in LAKS were about assessing writing, and they aimed to measure how knowledgeable the participants were in this domain. The findings revealed that the participants got a mean score of 6.573 over 15. As for the items in this section, the highest mean score belonged to the item “using visuals which guide learners for writing poses a problem (False)” which was answered correctly by most of the participants. On the other hand, “giving learners an opinion and asking them to discuss it is a valid way of assessing their writing skills (False)” received the lowest mean score, and was answered incorrectly by most of the participants.

The last section in LAKS focused on the participants’ knowledge level in assessing speaking, and the items between 46 and 60 aimed to measure how knowledgeable they are in assessing speaking. The mean score was close to the one in writing, and the participants demonstrated a mean score of 6,808 over 15, which again means that their assessment knowledge in speaking was less than half. In this skill, the item with the highest mean score was found to be item “giving learners one task is enough to assess speaking skills (False)” which was answered correctly by most of the participants. The lowest mean score, on the other hand, belonged to the item “reading aloud is a technique used to assess speaking skills (True)” which was answered correctly by few participants.

In addition to the general and skill-based mean scores, the following figure presents a general picture of the participants’ scores based on percentages.

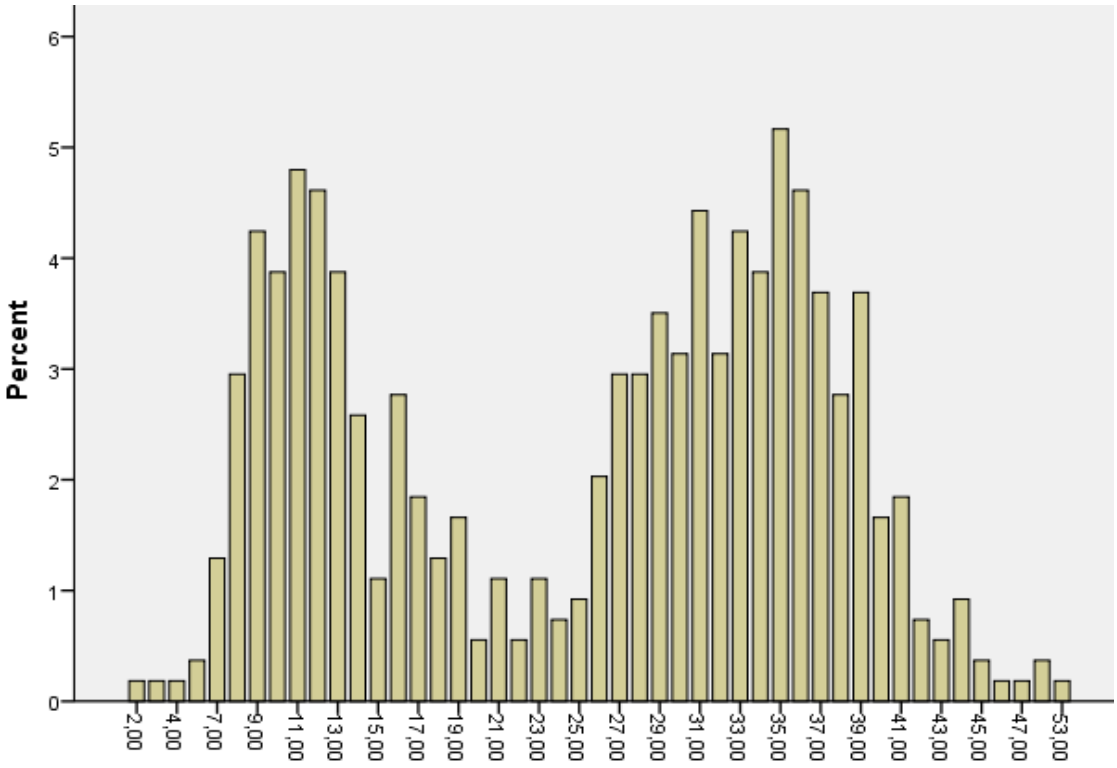

Figure 4.2. *The range of percentages based on the participants' scores*

When the figure is carefully examined, it can be seen that the lowest mean score was 2 out of 60 whereas the highest one was 53. The figure also shows that the range of the scores creates a kind of cut-off score on the figure which reveals that there were two groups one of which pulls the mean score to the lower (the participants with the mean scores between 7 and 20), and the other group pulling the mean score to a relatively higher level (the participants with the mean scores between 25 and 41). It is also seen that the number of the participants who got 45 and over from the scale (which means they have 75% and over correct answers) is quite few. Finally, the figure also makes it clear that there is a considerable number of participants who had poor performance in giving correct answers to the items, and also there are not many who can be regarded as high-performers in LAKS.

4.3. The Relationship among the Participants' Skill-based Assessment Knowledge

Another research question of the current study aimed to present how each skill-based knowledge correlated with the others and language assessment knowledge in

general. For the analysis, Pearson correlation was employed and the findings are presented in Table 4.7.

Table 4.8. *The relationship among skill-based language assessment knowledge*

	LAK	Reading	Listening	Writing	Speaking
LAK	1	,933**	,908**	,749**	,852**
Reading		1	,816**	,573**	,737**
Listening			1	,597**	,689**
Writing				1	,547**
Speaking					1

***Correlation is significant at 0.01 level; N=542*

The findings presented in the table demonstrated that all correlational values among the variables are significant. It was also found that all types of skill-based knowledge (assessing reading, assessing listening, assessing writing and assessing speaking) were highly and positively correlated with language assessment knowledge (LAK) in general. That means all types of skill-based assessment knowledge are important elements of LAK and if teachers are trained to be more knowledgeable in assessing one skill, it is highly probable that their LAK in general will increase as well. This finding might lead us to perceive language assessment knowledge as a holistic phenomenon with its own interrelated elements.

In addition to this, it was also revealed that all types of skill-based knowledge had high or moderate positive correlations among themselves. The highest correlational level was found between reading and listening (,816), whereas the lowest was between writing and speaking (,547) which is a moderate one. These high or moderate relationships among the skills mean that if EFL teachers' assessment knowledge in one skill increases, their assessment knowledge in others tends to increase in high or moderate levels. This finding again put forward that all types of skill-based assessment knowledge might be considered as interrelated elements.

4.4. Effects of Demographic Features on LAK Level of the Teachers

The fourth research question of the study examined the language assessment knowledge of the participants in terms of several variables such as years of experience,

educational background, the BA programme being graduated, workplace, testing course in BA, attending trainings on testing and being a testing office member. The findings belonging to each variable are presented in the tables below.

Table 4.9. *Language assessment knowledge according to years of experience*

years of experience	N	M
1-5 years	86	24,97
6-10 years	173	25,03
11-15 years	114	24,86
16-20 years	100	25,62
more than 21 years	69	25,75

	Sum of Squares	df	Mean Square	F	p
Between Groups	60,28	4	15,071	,115	.977
Within Groups	70129,14	537	130,594		
Total	70189,42	541			

The first variable was teaching experience and whether the participants' language assessment knowledge changed according to the years they spend in this profession was investigated. Among the participants, there are 86 teachers with 1-5 years, 173 teachers with 6-10 years of experience, 114 teachers with 11-15 years of experience, 100 teachers with 16-20 years, and 69 teachers with more than 21 years of teaching experience. To find the impact of teaching experience on LAK level of the participants, one-way ANOVA was used, and the findings revealed that there was no significant difference among the groups. Based on this, it can be said that teaching experience did not play a significant role on language teachers' LAK level.

Table 4.10. *Language assessment knowledge according to educational background*

Educational background	N	M
BA degree	238	25,508
MA degree	255	24,870
PhD degree	49	25,306

	Sum of Squares	df	Mean Square	F	p
Between Groups	50,805	2	25,403	,195	.823
Within Groups	70138,621	539	130,127		
Total	70189,426	541			

The second variable being the focus of that research question was the educational background. Among the participants, 238 teachers had BA degrees, 255 had MAs and 49 teachers had PhDs. To identify whether there was a difference among the groups, one-way ANOVA was used, and the results showed that the difference among the groups was not significant and the teachers' LAK level did not change according to their educational background.

Table 4.11. *Language assessment knowledge according to the programme being graduated*

BA Graduation	N	Mean	Std. Deviation	Std. Error Mean
English Language Teaching	347	25,42	11,722	,629
Non-ELT	195	24,76	10,790	,772

Mean diff.	df	t	p
,657	540	,644	.52

The third variable being examined in terms of LAK level was the BA programme from which the participants graduated. There were 347 teachers graduating from English language teaching departments whereas 195 teachers were the graduates of non-ELT departments such as English language and literature, English linguistics or translation and interpretation. In order to identify the role of BA programmes on LAK level, independent samples t-test was utilized, and it was found that there was not a statistically significant difference between ELT and non-ELT graduates in terms of their LAK level. In other words, the programme being graduated, whether ELT or non-ELT, did not influence the language assessment knowledge of the teachers.

Table 4.12. *Language assessment knowledge according to the workplace*

Workplace	N	Mean	Std. Deviation	Std. Error Mean
at a state university	372	25,346	11,565	,599
at a private university	170	24,847	11,022	,845

Mean diff.	df	t	p
,499	540	,474	.63

Whether the teachers in this study worked at a state or private university was another variable investigated in this research question. The number of the teachers working at a state university was 372 and on the other side, 170 teachers worked at a private university. Based on the results obtained through the independent samples t-test, it can be seen that there was not a significant difference between these two groups and workplace was found to have no effect on the teachers' language assessment knowledge.

Table 4.13. *Language assessment knowledge according to testing course in BA*

A separate testing course in BA	N	Mean	Std. Deviation	Std. Error Mean
Yes	260	25,019	12,045	,747
No	282	25,347	10,769	,641

Mean diff.	df	t	p
-,328	540	-,335	.73

The effect of the testing course participants had during their BA programme was also investigated as a variable. Among the participants, 260 teachers responded that they had had a testing course in their BA programme while 282 teachers stated that they had not taken a course on testing and assessment. To see the difference between the participants who had taken a course on testing and assessment and who had not, independent samples t-test was applied, and the findings revealed that there is not a significant difference between these two groups. In other words, it can be said that the testing and assessment course given in BA programmes had no effect on the teachers' LAK level.

Table 4.14. *Language assessment knowledge according to the attendance to trainings*

Attending any trainings in LTA	N	Mean	Std. Deviation	Std. Error Mean
Yes	282	25,741	11,967	,712
No	260	24,592	10,720	,664

Mean diff.	df	t	p
1,148	540	1,17	.24

Another variable focused was the attendance of the participants to professional development activities on testing and assessment, and they were asked to respond whether they attended any training/courses or seminars. Among them, 282 participants stated that they had attended such trainings while 260 of the total number expressed no attendance to such activities. The results derived from the independent samples t-test analysis indicated that there was no significant difference between these two groups, which gives the conclusion that the training received on language assessment did not have a significant impact on the teachers' LAK level.

Table 4.15. *Language assessment knowledge according to being a testing office member*

Being a testing office member	N	Mean	Std. Deviation	Std. Error Mean
Yes	260	26,303	11,710	,726
No	282	24,163	11,007	,655

Mean diff.	df	t	p
2,140	540	2,19	.02

Among all variables examined in this research question, the only significant difference was found in terms of being a testing office member or not. The number of the participants who had been a testing office member at their university was 260 whereas 282 participants were not the members of the testing office in their institution. The mean score of testing members was 26,30 whereas the score of non-testing members was found to be 24,16. Independent samples t-test was administered, and the findings showed that though the mean scores were slightly different, there was a statistically significant difference between these two groups, and the LAK level of the participants having worked as a member of testing office was higher than the others. Based on this, it can be concluded that working on testing, doing institutional staff and being involved with some practical elements related with testing and assessment might have a positive impact on LAK level of the teachers.

4.5. Perceived Self-competency and Actual Language Assessment Knowledge Level

Perceived self-competency of the teachers in language assessment knowledge was another research matter of the current study, and whether their LAK level changed

according to their perceived self-competency was investigated based on each language skill. The competency variables were initially coded as very competent, competent, not very competent and not competent. However, after the data were collected, it was seen that the number of the participants choosing "not competent" was very few in all skills (from two to seven participants); so, these participants were combined with "not very competent" category before the final analysis. The findings derived from one-way ANOVA analysis are presented in the following tables.

Table 4.16. *Perceived self-competency of the teachers and their LAK level in assessing reading*

Assessing Reading	**N**	**M**
very competent	152	6,769
competent	355	7,019
not very competent	34	8,676

	Sum of Squares	**df**	**Mean Square**	**F**	**p**
Between Groups	102,202	2	51,10	2,567	.078
Within Groups	10709,244	538	19,90		
Total	10811,445	540			

Table 4.15 above gives the findings related with the reading skill in which the teachers demonstrated a relatively higher level of LAK (7,055 over 15) compared to other skills. Although, their ratio of success in assessing reading was found to be less than 50% in general, the findings above showed that almost 95% of the participants perceived themselves competent or very competent. On the other hand, the ones who thought that they were not very competent in assessing reading, had the highest mean score among all. According to the findings, no significant difference was found among the participants who perceived themselves as very competent, competent, and not very competent in terms of their LAK level in reading. However, it can be clearly seen that the teachers' perceived self-competency in assessing reading is far from their actual LAK level.

Table 4.17. *Perceived self-competency of the teachers and their LAK level in assessing listening*

Assessing Listening	N	M
very competent	112	4,821
competent	338	4,695
not very competent	89	4,943

	Sum of Squares	df	Mean Square	F	p
Between Groups	4,843	2	2,42	,222	.80
Within Groups	5834,760	536	10,88		
Total	5839,603	538			

In the findings of the previous research questions, the participant teachers demonstrated the lowest LAK level in listening (4.752 over 15) among all skills, which means that they were the least knowledgeable in assessing listening. On the contrary, the findings regarding their perceived self-competency tell the opposite since more than 80% of the teachers perceived themselves as competent or very competent. In addition to this, there was not a significant difference among the perception groups in terms of their LAK level in assessing listening, and again, it was found that the ones who perceived themselves as not very competent had the highest mean score compared to the others.

Table 4.18. *Perceived self-competency of the teachers and their LAK level in assessing writing*

Assessing Writing	N	M
very competent	161	6,347
competent	333	6,657
not very competent	45	6,866

	Sum of Squares	df	Mean Square	F	p
Between Groups	14,381	2	7,19	1,181	.30
Within Groups	3264,695	536	6,09		
Total	3279,076	538			

In terms of their LAK level in assessing writing, no significant difference was found among the perception groups. It was also revealed that more than 90% of the teachers perceived themselves as competent or very competent in assessing writing though their actual LAK level in this skill was 6.573 over 15. This finding again reveals a huge gap between the participants' perceived self-competency and their actual level. Finally, it is

again seen that the ones who perceived themselves as not very competent had the highest mean score compared to the other groups.

Table 4.19. *Perceived self-competency of the teachers and their LAK level in assessing speaking*

Assessing Speaking	N	M
very competent	129	7,0620
competent	336	6,6786
not very competent	74	6,918

	Sum of Squares	df	Mean Square	F	p
Between Groups	14,851	2	7,42	,953	.38
Within Groups	4174,303	536	7,78		
Total	4189,154	538			

For the last skill, assessing speaking, the findings were similar to the others. There was no significant difference in terms of LAK level in assessing speaking among the participants based on their perceived self-competency in this skill. Again, almost 85% of the teachers perceived themselves as competent or very competent though they demonstrated a LAK level of 6.808 over 15, which shows a difference between their perceptions and actual level. Finally, the last important point was again similar to the other skills and the ones with "not very competent" perception had a relatively higher level of LAK in assessing speaking compared to the other groups.

4.6. The Opinions of EFL Teachers Regarding Their LAK Level and the Findings of the Scale

The current study employed a mixed-method design based on a QUAN ➔ qual sequence which had qualitative elements following the quantitative data collection and analysis in order to have an in-depth understanding of the quantitative findings. The sixth research question aimed to investigate the participants' opinions regarding their general and skill-based LAK level and the findings derived from the statistical analysis. The teachers participated in the qualitative phase were asked to write detailed answers for five questions in the open-ended protocol focusing on the findings related with the findings of the quantitative data of the participants. Their answers were analyzed based on the

qualitative content analysis scheme of Creswell (2012) and the findings related with these questions are presented in the following table.

Table 4.20. *Analysis of the qualitative data - 1*

QUESTION	CODES	THEMES
Q-1.	a) Limited exposure in the curriculum (x3) b) Teacher educators' insufficient knowledge (x2) c) Non-ELT graduates (x2)	1. Lack of knowledge in LTA 1. A. Insufficiency of pre-service education
	a) Insufficient professional development activities (x4) b) Lack of motivation of teachers (x4) c) Lack of sources in LTA (x2)	1.B. Insufficiency of in-service education
Q-2.	a) Teaching reading is a priority in the curriculum (x4) b) More concrete outcomes in reading (x2) c) More experience in teaching and assessing (x2) d) More resources for this skill (x2)	1. It is easy to teach and assess reading
	a) Due to practicality issues (x3) b) Teaching listening is not a priority (x2) c) Insufficient experience in assessing listening (x2)	2. Assessing listening is challenging
Q-3.		
	a) Assessment not related with demographics (x2) b) Lack of knowledge due to the reasons in the first question (x1)	1. No impact of demographics on LAK
Q-4		
	a) Feeling the need to improve themselves (x4) b)Testing members attending trainings and conducting research (x4) b) Practice opportunities in testing office (x3) c) Non-members being far from LTA (x1)	1. The more you are involved, the more you learn
Q-5		
	a) Being unaware of their assessment knowledge level (x6) b) Resistance to accept their incompetency (x3) c) Being unaware of the importance of LTA (x2)	1. Self-perception not reflecting the reality

The findings derived from the participants' answers are listed in the table above as codes and emerging themes. According to this, it was revealed that the main reasons of the lack of knowledge in LTA among the teachers in Turkish higher education setting was the insufficiency of education and training in both pre-service and in-service levels. As for the pre-service level, the teachers stated that limited exposure in the curriculum, teacher educators' insufficient knowledge and non-ELT graduates were the major reasons of the insufficiency of knowledge. On the other hand, insufficient professional development activities, lack of motivation among the teachers and lack of sources in language testing and assessment were uttered as the main reasons of their insufficient knowledge during in-service level.

The second question focused on the participants' opinions related with skill-based findings of the quantitiative data which put forward that the teachers were the most knowledgeable in assessing reading and the least in assessing listening. Related with assessing reading, the participants expressed in open-ended questions that the teachers were more knowledgeable in assessing reading because it was regarded as an easy skill to teach and test due to its priority in the curriculum, its concrete outcomes, teachers' having more experience in teaching and testing it, and having more resources for this skill. In addition to this, the participants also commented on assessing listening, found to be the skill in which the teachers were the least knowledgeable, and the major reason for this was the perception of listening as a challenging skill due to several practicality issues related with teaching and testing it, its not having priority in teaching and insufficient experience in assessing listening.

The quantitative data of the study revealed that six of the seven demographic features had no effect on LAK level of participants and the third question in the open-ended protocal asked the participants to comment on this. On this issue, they believed that language assessment knowledge was not a phenomenon related with demographic features of people and they expressed similar reasons to the ones in the first question. Furthermore, the fourth question focused on the only demographic feature, being a testing member or not, which significantly influenced LAK level of the teachers and the participants were asked about the underlying reasons of it. The theme emerged from their answers to this question was "the more you are involved, the more you learn", and the major reasons for this thought were testing members' feeling the need to improve themselves, their attending to trainings and conducting research, practice opportunities in testing office and non-members' being far from LTA.

The last open-ended question within the scope of the sixth research question was about the difference between the teachers' perceived self-competency and their actual LAK level. It was derived from the quantitative data that though the teachers perceived themselves very competent or competent in assessing all skills, their actual LAK was not at the same level with their perception, and the participants were asked to comment on this in open-ended questions. Their answers put forward that being unaware of their assessment knowledge level, resistance to accept their incompetency and being unaware of the importance of LTA were the main reasons of self-perception's not reflecting the reality in terms of their LAK level.

4.7. EFL Teachers' Needs in Language Testing and Assessment

Based on the last two questions in the open-ended protocol, the last research question of the study aimed to explore the participants' opinions regarding their needs in assessing four language skills and the ideal features of an in-service training module on language testing and assessment they want to have. The codes and emerging themes derived from the participants' answers to the last two questions are presented in the table below.

Table 4.21. *Analysis of the qualitative data - 2*

QUESTIONS	CODES	THEMES
Q-6.	a) Trainings and workshops for all skills (x4) b) Overcoming subjectivity in productive skills (x4) c) Constructing tests/tasks for assessing each skill (x3) d) Analysing the validity and reliability of tests (x2)	1. Needs in assessing four skills
Q-7.	a) Given by professional LTA practitioners (x6) b) Hands-on practices in trainings (x3) c) Both theory and practice (x3) d) Long-lasting and sustainable (x2) e) Institutional factors considered (x2)	1. What kind of a training module

The sixth question in the open-ended protocol asked the participants to express their opinions regarding their needs in assessing four skills. The findings for this question revealed that trainings/workshops for assessing each skill, how to overcome subjectivity in scoring productive skills, constructing tests/tasks for assessing each skill and analyzing the validity and reliability of the tests were at most important as their needs to be more knowledgeable in assessing all skills.

Finally, the last open-ended question focused on the participants' thoughts regarding the features of a training module on language testing and assessment. It was found that the participants highly emphasized the trainers' being professional language testing and assessment practitioners. Besides, it was expressed that the training sessions should not only include theoretical knowledge in LTA but also hands-on practices on how to assess learners in all skills. The last points touched upon by the participants in their answers were the facts that these trainings should be long-lasting and sustainable, and institutional factors should be taken into consideration while preparing the content of the trainings.

5. DISCUSSION

5.1. Psychometric Properties of LAKS

In the literature, research focusing on the assessment literacy of teachers generally employed certain measurement tools, and revealed the assessment level of teachers in pre-service and in-service (Impara, Plake & Fager, 1993; Campell, Murphy & Holt, 2002; Mertler & Campell, 2005; Malone, 2013). However, specifically in language assessment literacy which "is still in its infancy" (Fulcher, 2012, p. 117), there is not a validated tool measuring language assessment literacy of teachers. The lack of such an instrument has been the major problem in determining language teachers' assessment knowledge level, which is regarded as the core of language assessment literacy.

Based on this background, this study aimed to develop Language Assessment Knowledge Scale – LAKS. After a thorough validation process which included literature review, meetings with language teachers and testing and assessment practitioners, expert opinion and a piloting process, LAKS with 60 items and four constructs (assessing reading, assessing listening, assessing writing, and assessing speaking) was completed by 542 EFL teachers working at higher education context. The findings derived from second order confirmatory factor analysis revealed a perfect model-data fit. Though some of the items in assessing writing and assessing speaking constructs had low factor loadings and item total correlations, they were decided to be kept in the scale based on the expert opinion considering their significant contribution to the content validity of the scale. Besides, the scale demonstrated satisfactory levels for reliability according to the Cronbach alpha analysis. Based on all the validation and statistical procedures, it can be concluded that LAKS can be used as a valid and reliable instrument to measure language teachers' language assessment knowledge.

The major underlying reason behind developing a measurement tool on language assessment knowledge was the urgent need in this field. In other words, in addition to its statistical and procedural validation, LAKS not only presents a baseline for researchers interested in this field but also creates an initial framework to understand language assessment knowledge level of EFL teachers. Besides, this scale might be used not only as the first step to identify the needs of in-service teachers and to develop training programmes, it might also have implications for pre-service teacher training programmes in preparing future teachers more equipped for the field.

5.2. Language Assessment Knowledge Level of EFL Teachers

The second research question investigated general and skill-based Language Assessment Knowledge (LAK) level of EFL teachers in Turkish higher education setting. The mean score is 25 out of 60, indicating that, on average, more than half of the items in the scale were answered incorrectly by the participant teachers. In other words, the teachers in the current study indicated a relatively low level of language assessment knowledge.

In the literature, there are some studies focusing on the assessment knowledge of teachers especially in general education; however, in ELT, there are few studies aiming at revealing the language assessment knowledge of teachers. The rarity of these studies in ELT was also mentioned by Hatipoğlu (2017). Though limited in number, these studies also support the findings of the present study. To start with, Tao (2014) developed and validated Classroom Assessment Knowledge Test which is composed of multiple choice items designed to measure assessment knowledge base of 104 in-service EFL teachers in Cambodia. It was found that the participants had limited assessment knowledge. The next study belongs to Mede and Atay (2017). 350 teachers took part in their study, and the data were collected via an online LTA questionnaire that was adapted from Vogt and Tsagari (2004). By using this questionnaire, they aimed to reveal classroom-oriented LTA practices and needs of Turkish EFL teachers. The results suggested that the respondents had limited assessment knowledge, and they were not good at testing four skills. Additionally, Xu and Brown (2017) conducted a study with 891 English teachers who work in China. To obtain data from the participants, they made use of the adapted version of Teacher Assessment Literacy Questionnaire. It was found that most of the teachers had either a very basic or minimum level of assessment literacy. Popham (2009) also stated that most of the teachers do not have adequate knowledge related to language assessment, and discussed the severity of the situation by saying that for most of the teachers, test "is a four-letter word, both literally and figuratively" (p. 9). All of these studies mentioned above show parallelism with the findings of the current study revealing this relatively low assessment knowledge of language teachers.

There might be many reasons leading to that relatively low mean score, and these reasons were among the foci of the qualitative data. The first open-ended question in the qualitative phase asked the participants to comment on the LAK level (25 out of 60) of language teachers in Turkish higher education setting. The answers demonstrated that the

main reason for this was the insufficiency of education on language testing and assessment (LTA) in pre-service and in-service levels. As for the pre-service education, the participants mainly focused on the limited number of courses in pre-service programmes and limited exposure to knowledge in LTA. This finding is in line with the study conducted by Hatipoğlu (2015). Besides, Herrera and Macias (2015) also stated that LAL should not be restricted to one course in pre-service education. For this issue, the following expressions of the teachers in the present study serve as a summary of the participants' thoughts.

> *"The main reason is that we had just one course on testing and assessment in our pre-service programme and in that course, we covered general topics such as validity, reliability, washback etc. We received very limited knowledge on assessing language and that is why we have a relatively low level."*

Another important point related with the insufficiency of pre-service programmes was touched upon by two participants, and it was seen that the competency of teacher educators giving LTA courses were not found at a desired level.

> *"At those times, she was teaching testing and assessment superficially. We had a book, the course was totally based on that book, I mean, a little bit far away from real life. When I started working and I met with real testing and assessment practices, I was sure that course would have been more effective to prepare us for our future profession. What she taught was quite different from the reality."*

This finding is in parallel with Stiggins (1999) and Hatipoğlu (2015) who believe that the teacher educators who give language testing and assessment courses at university should have a lot of knowledge regarding language assessment. Jeong (2013) also stated that professional background of teacher educators giving these courses is important.

Two respondents also expressed their ideas regarding the insufficiency of pre-service education related with non-ELT graduates. They stated that for non-ELT graduates who did not even receive such a course during their undergraduate education, this was a bigger problem. For this, one teacher uttered that:

> *"Even ELT graduates have difficulty in language testing and assessment though they have one course on this subject in pre-service education. It is nearly impossible for non-ELT graduates to have enough knowledge related to language testing and assessment."*

The other important reason in the eyes of the participants related with the low knowledge level was the insufficiency of in-service education. They believed that teachers in higher education setting did not receive enough training on LTA, and for this

reason, their knowledge is limited. According to them, the most critical point here is lack of professional development activities, and this finding shows parallelism with Köksal (2004) and Lam (2015)'s studies in which lack of sufficient training was stressed. Besides, Mendoza and Arandia (2009) stated that there should be more training for language teachers in language assessment. At this point, the sentences written by one of the participants below illustrate the situation very well.

> *"After we graduated and started to work, we had very few or no opportunities to improve ourselves in testing and assessment. Some private universities provide such opportunities but at state universities, we do not have this chance. Thus, it is hard to improve and keep yourself updated if you do not have these opportunities."*

In addition to the lack of opportunities, the participants also mentioned that LTA as a subject matter was not attractive for them, and most of the teachers were not motivated enough to be more knowledgeable in LTA. The expressions below provide a good example to explain this situation.

> *"To be honest, it is a difficult topic for most of us and we do not feel enthusiastic to improve ourselves. Maybe we think that it is the duty of people in testing office and not ours."*

Besides, lack of sources in LTA was mentioned by two participants. They stated that there are not enough sources to improve themselves in LTA. The following utterance is an example for their opinion:

> *"Although there are many books focusing on language testing and assessment, there are very few which are solely based on how to assess each skill. Additionally, these books are not some of the books that are available in libraries; thus, it is difficult for us to get these books."*

As is seen, according to the participants, the main factor associated with the lack of knowledge in LTA is highly related with insufficiency of education in pre-service and in-service levels. In Turkey, ELT programmes at universities include just one course on testing and evaluation covering the elements on a surface level in their programme because of the time constraint since the content of this course has to be covered in one academic term period. For this reason, pre-service teachers graduate without receiving enough theoretical and practical opportunities in this field. In their in-service years, they hardly have opportunities to improve themselves in this topic as well. In other words, the combination of these two insufficiencies makes EFL teachers in Turkish higher education context have relatively low level of LTA.

5.2.1. Skill-based language assessment knowledge of EFL teachers

The other focus of this research question was to find out the language assessment knowledge level of the participant teachers in respect to each skill that are reading, listening, writing and speaking. The discussion in this part starts with the order of the skills based on their mean scores, and goes on with the discussion of each and every item in each skill in relation to the literature.

The findings indicated that the highest mean score belongs to assessing reading whereas the lowest mean score belongs to assessing listening. It is obvious that the participant teachers are more competent and knowledgeable in assessing reading when compared to assessing other skills. Why assessing reading has higher mean scores can be found in the utterances of Hubley (2012), and Backlund, Brown, Gurry and Jandt (1980). Hubley (2012) stated that there is agreement among scholars in the argument that reading is a crucial skill, and even maybe the most important one, and much of the input comes from reading sources surrounding us. Because of the density of input surrounding the learners in the classroom as well, learners have to read a lot. As reading skill is given importance, teaching it is highly valued, and, it is assessed by the teachers as a natural consequence. There are various ready-made materials for assessing reading; thus, it does not become a challenge for teachers to assess reading skills of their learners (Backlund, Brown, Gurry, & Jandt, 1980). Why the other three skills had lower mean scores were mentioned in the literature by touching upon the difficulties each skill possesses.

For assessing writing, Weigle (2012) expressed that assessing writing could be perceived as something easy, and people may think that teachers only give the topic and ask learners to write on that topic. Indeed, it is not as easy as people may think, because just giving the topic and asking learners to write on that topic are not a good way of assessing writing. Weigle (2012) touched upon the problems of assessing writing, and demonstrated that assessing writing is not an easy task. Speaking is also regarded highly important because of the oral communication taking place a lot in our lives (Heaton, 1990). Madsen (1983) stated that speaking is the most difficult skill to assess because of its subjectivity and complex nature, because teachers do not know what and how to assess regarding speaking skill. In other skills, they have ready-made materials provided by the coursebooks and publishing companies; but, in speaking they are all alone.

As for listening, Flowerdew and Miller (2012) discussed that assessing listening is perceived by both learners and teachers as an issue which somehow improves by itself.

Buck (2001) also mentioned this problem by saying that listening is neglected in terms of teaching and assessing, which is one of the findings of the current study. To draw attention to the ignorance, Flowerdew (1994), Nunan and Miller (1995) and Flowerdew and Miller (2005) stated that listening skill is a 'Cinderella' skill which majority of teachers take for granted. For Buck (2001), why listening is neglected lies on the complicated nature of listening as a skill and practicality issues related to assessing listening.

The qualitative data of the study also focused on the knowledge level of the teachers in assessing reading and assessing listening, former being the highest and latter being the lowest. As for assessing reading, the participants thought that it was a skill that is easy to teach and test in general, and for this reason, the teachers had the greatest knowledge in it. The major factors that made the teachers think that reading was easier to test were related to giving priority to this skill with more concrete outcomes in the curriculum and the accessibility of more resources in assessing reading. One of the respondents wrote the following sentences to express her ideas on this issue.

> *"It is a dominant skill. You teach vocabulary and grammar through reading activities and it is an indispensable part of our classroom teaching... Also what you want to teach is quite clear in reading, and I can say it is easier to teach it compared to other skills. That might be a reason."*

Another important factor that made the teachers perceive reading as a skill easier to teach and assess was found to be the experience of the teachers in this skill. The following expressions summarize this experience very well:

> *"As a learner, I was used to reading classes. We were given reading texts, and I have some experiences rooted in my high school years. As a teacher, coursebooks are also full of examples intended for teaching and assessing reading; so, I feel myself more experienced in teaching and assessing reading".*

Another important factor touched upon by the participants was the existence of more resources for reading. They believed that the existence of more resources makes teaching and assessing reading easier, and this is reflected in the assessment knowledge level of the teachers. The following sentences written by two of the participants explain this factor well.

> *"Teachers feel more comfortable because there are a lot of sources both to teach and test reading, and even with the guidance on how to use them."*

Considering the mean scores of each skill in the scale, assessing listening got the lowest mean score. The opinions of the participants were asked regarding the possible

reasons for this low mean score. The respondents stated that assessing listening is challenging for most of the language teachers. One reason was found as the practicality issues in assessing this skill. One of the teachers uttered the following sentences on this issue.

> *"We cannot modify the materials used in assessing listening. Adding extra sentences, or cutting some parts of a listening material are really demanding. We even do not know how to do this, because doing this requires extra competencies apart from ELT knowledge. Because of this, we tend to make use of ready-made materials which in turn makes us not question the appropriateness of the existing materials."*

Another possible reason of this low mean score was expressed by the teachers as teaching listening was not a priority for them. According to them, listening was the least favoured skill among all, and it was even a problematic skill for the teachers themselves. The sentences below expressed by one of the teachers are quite striking.

> *"I do not feel myself competent enough in teaching listening. Naturally, I cannot assess a skill efficiently in which I have difficulty in teaching."*

Another participant shared similar sentences as follows.

> *"There are not enough activities in the class to teach listening to our learners. We cannot find materials suitable for our learners in terms of content or vocabulary. Most of the language teachers have weaknesses in listening; so, it becomes a challenge for them to assess this skill."*

The data derived from the second question focusing on the participants' ideas on findings related with assessing reading and listening put forward that they perceived these findings as quite natural and parallel with their real life experiences. They strongly believed that their knowledge and experience in teaching these skills were hand in hand with their knowledge in assessing them. It was also seen in detail that as for the listening skill, they not only felt themselves incompetent in teaching listening but also assessing it. For this reason, it can be concluded that teachers may need training in both teaching and assessing this skill.

What has been discussed so far under the second research question is mainly related to the teachers' knowledge level in assessing reading skill, ranked as the highest in terms of mean scores, and knowledge level in assessing listening skill, ranked as the lowest, and writing and speaking between them. The next is the discussion of each and every item in separate skills in relation to the literature.

5.2.1.1. Assessing reading

The item which has the highest mean score in assessing reading is "one reading text is enough to be included in a reading exam". The answer to this item is false, because one reading text is not enough to be included in assessing reading skills of learners. Alderson (2000) mentioned that many reading tests have short reading passages to assess reading skills of learners, and TOEFL is an example for these tests. When as many texts as possible are included in a reading exam, learners are given new chances for new starts which eventually results in increased reliability (Harris, 1969; Hughes, 1989; Douglas, 2010, Green, 2014). In the current study, 73% of the participants, 400 teachers, answered this item correctly; so, this means that most of the teachers are aware of the fact that they should include at least two tasks in a reading exam. In other words, most of the teachers have this knowledge. Apart from 400 teachers answering this item correctly, there are 108 teachers who gave an incorrect answer to this item, and 34 teachers selected don't know option. What can be concluded from this finding is that 108 teachers know this item incorrectly whereas 34 teachers are aware of the fact that they do not have this knowledge.

"Reading texts in a reading exam include various genres (essay, article, etc.)" has the second highest mean score of all the items in assessing reading. Alderson (2000), Douglas (2010) and Hubley (2012) stated that learners should be exposed to various genres and formats from books to short official announcements. The reason behind this is that each and every genre has its unique features, and is different in length; thus, the learners get used to various use of language in different genres. Hughes (1989) also discussed the washback effect of testing on learning. If learners are not presented various genres in reading exams, then they will tend to read limited range of reading texts. 60% of the participant teachers, that is 328 teachers, gave a correct answer to this item. It can be concluded that more than half of the participant teachers know that they should make use of various genres in a reading exam to make their learners familiar with different kinds of reading texts. On the contrary, there are 188 teachers who answered this item incorrectly, and 26 teachers who stated that they do not know the answer of this item. The number of the participants choosing don't know option is the second lowest one in number in assessing reading part.

The third highest mean score belongs to the item "using ungrammatical distractors in multiple choice questions in a reading exam is a problem". The answer to this item is true. Heaton (1990), Alderson (2000) and Purpura (2004) stated that test designers tend

to neglect grammatical appropriacy of each and every option in multiple choice questions; however, these ungrammatical distractors could be identified by some of the learners immediately just because of the fact that they are ungrammatical. The purpose in assessing reading is whether learners have understood the text or not, or whether they can identify the option which is not mentioned in the text or is irrelevant, etc. Hence, using ungrammatical distractors does not serve its purpose. One more reason could be that learners are given wrong input in terms of language, which is not desired. In well-designed multiple choice tests, all of the options and the distractors should be grammatically correct (Madsen, 1983; Osterlind, 1989; Salend, 2009; Hubley, 2012). 54% of the respondents (N=296) stated that all the distractors in the options should be grammatically correct. Among the rest, 47 respondents selected don't know option. It is clear that more than half of the participants have this knowledge in themselves, and they know that they should not use ungrammatical distractors in multiple choice questions while designing tasks in assessing reading. Though high in mean score when compared to the other items in assessing reading, this item still does not have a very high mean score. The reason behind this could be that there are some tasks and procedures used a lot in assessing reading skills, and asking questions in multiple choice format is one of them. Indeed, coming up with ungrammatical distractors might be easier for the teachers since they encounter lots of samples from their students in practice. By using these ungrammatical distractors, the teachers might also be thinking of getting their students' attention to these mistakes with their options in the choices.

"Taking vocabulary difficulty into consideration is necessary in assessing reading skills" has the fourth highest mean score. Madsen (1983), Heaton (1990), Alderson (2000), Douglas (2010) and Read (2012) stated that vocabulary plays a great role in reading. As the role of vocabulary is great, then the selection of the vocabulary items to be used in a reading text should be given great care, because vocabulary knowledge and reading comprehension are interrelated (Perfetti & Adlof, 2012; McKenna & Stahl, 2015). Furthermore, vocabulary is the key to the understanding of the reading text, and also successful completion of the test. Hubley (2012) suggested that the number of the unfamiliar words should be limited in a reading text, around 5-10%. 53% of the teachers answered this item correctly, which is slightly more than the half. The number of the teachers giving a correct answer is 288 whereas 224 participants answered it incorrectly. The numbers of the teachers giving a correct and an incorrect answer to this item are so

close to each other. Only 30 teachers selected don't know option for this item. These results suggest that nearly half of the teachers know and aware of the relationship between the selection of vocabulary and reading comprehension.

"Asking learners to summarize the reading text is a way of assessing their reading skills" is another item that has one of the highest mean scores. Learners are asked to summarize the main points and key information mentioned in the reading text which is a way of assessing their reading skills (Alderson, 2000; Douglas, 2010; Hubley, 2012; Perfetti & Adlof, 2012). Via summarizing, teachers have the chance to check whether learners have understood the reading text or not, because there is not a chance option here as in multiple choice test. If learners have understood, then they can summarize the reading text (Brown, 2003). 49% of the teachers responded to this item correctly, that is nearly half of the participants. 269 teachers gave a correct answer to this item whereas 257 teachers gave an incorrect answer to it. The number of the teachers who selected don't know option is 16, which is the lowest number in don't know option in assessing reading. It is obvious that half of the participants know that reading skills could be assessed via summarizing, and at the same time half of them do not have this knowledge. The ones who do not have this knowledge may have relied on their experiences while answering the scale, and if summarizing is not a task used in assessing reading in their institutions, they may not be familiar with it.

"In top-down approach, assessment is on overall comprehension of the reading text" is an item whose answer is true. Alderson (2000) and Hubley (2012) stated that top-down approach was suggested by some psycholinguistic theorists such as Goodman (1967), and in this approach, learners are actively involved in larger bits of information in the reading text, and they have to care about the meaning of the text. Learners are expected to understand the overall meaning of the reading text, and they have to identify the main ideas and supporting details. Thus, the meaning of the reading text is of great concern in top-down approach (Carrell, 1998; Eskey & Grabe, 1998; Brown, 2003). 49% of the teachers answered this item correctly (N=267), and the mean score is the same with the previous item. Apart from these 267 teachers answering it correctly, 110 of them answered it incorrectly. The number of the respondents choosing don't know option is the highest in assessing reading with 165 teachers. The reason why the number of the teachers choosing don't know option is too high could result from the terminology used in this item. Maybe the teachers do not know what top-down approach is. The results of

this item indicated that nearly half of the participants know top-down approach and the purpose of it while assessing reading skills.

The next item is "the language of the questions is simpler than the text itself" the answer to which is true. Madsen (1983), Hughes (1989), Osterlind (1989) and Hubley (2012) uttered that the level of the questions should be less difficult than the reading text. The reason behind this was discussed by Heaton (1990) as follows. If a simple reading text is followed by more difficult questions, then the learners have difficulty understanding the questions or statements related to the reading text. However, the purpose is to check whether learners are able to understand the reading text or not. In this case, even though the learners could understand the text, because they cannot comprehend the questions due to their high level of language, learners cannot answer the questions correctly. Additionally, Alderson (2000) stated that the reason of learner's poor performance cannot be determined in such cases. Whether it is because of the difficult questions or because of the difficult reading text cannot be identified fully. Owing to this fact, the language of the questions should be much simpler than the reading text itself. 48% of the respondents gave a correct answer to this item, which is slightly lower than half of the total number. In this item, the number of the teachers answering this tem correctly (N=264) and incorrectly (N=220) are close to each other, implying that there are not many teachers who selected don't know option (N=58). The findings displayed that half of the participants are aware of the ideal language level of the questions, and the link between the questions and the reading text in terms of language level.

"Cloze test is used for assessing the main idea of the text" is another item which does not have one of the highest or lowest mean scores in reading part. The answer to this item is false. Heaton (1990) drew attention to the difference between cloze tests and gap-filling tests. In gap-filling tests, the deletion is arbitrary, that is, there is not a systematic rule in the deletion of the words. However, in cloze test, there is systematicity in terms of deletion. For instance, every *nth* word is deleted, and the decision of every *nth* word that is going to be deleted is determined by the teacher as the starting point. After the first decision, the teacher has no control over the words and phrases. Alderson (2000) uttered that as cloze test is mostly word-based, there is no opportunity to assess reading skills of learners, and added that most deleted words in cloze tests can be answered correctly by just having a look at the previous or forthcoming two or three words, and there is no need to understand the long discourses or main ideas. In other words, learners have to process

the grammar and vocabulary at the sentence level (Cash & Schumm, 2006; McKenna & Stahl, 2015), and it has nothing to do with the main idea of the text. In addition to this, as the teacher has no control over the deleted items, what is to be tested including the main idea of the text cannot be determined here. 46% of the participants' answer is correct; however, there are more teachers who answered this item incorrectly or chose don't know option. Even though there are 250 teachers answering this item correctly, 230 of them gave an incorrect answer to it, and 62 of them were aware of the fact that they do not have this knowledge related to the item; hence, they selected don't know option.

"Using the same words in the correct option as in the text is not a problem" is an item which is ranked nearly in the middle of the 15 items in terms of mean score in assessing reading. Heaton (1990) stated that multiple choice tests are one of the most commonly used tests for assessing reading, but a great care should be taken while using these types of questions. Effectiveness of each item used is an important issue; however, if the same words are used in the reading text and in the options, then effectiveness of each item decreases (Harris, 1969; Madsen, 1983; Osterlind, 1989; Purpura, 2004; Salend, 2009). 44% of the participants (N=243) answered this item correctly, which is less than half of the total number. The number of the teachers who answered this item incorrectly are nearly the same as the ones giving an incorrect answer (N=241). There were 58 participants who were not decisive about the item being true or false. It is obvious that not many respondents know the effective use of multiple questions following reading texts. If the same words are present in both the reading text and the option, then the question is not evaluative at all. It is sufficient for the learners to identify the recurrent words, and they do not have to understand the reading text and the questions to have a correct answer. As the aim is to assess reading skills of learners, this type of question does not serve its purpose.

"Errors of spelling are penalized while scoring" is the next item. Hughes (1989) drew attention to the issue by saying that errors of grammar and spelling should not be penalized, because the main goal is to test reading, not something else. Furthermore, Hughes (1989) and Heaton (1990) added that if grammar or spelling, that are other purposes, are integrated into the reading text, then the reliability of the reading text decreases. Additionally, Cash and Schumm (2006) stated that spelling errors should be allowed, because "spelling is a production task" (p. 119). 43% of the teachers (N=237) answered this item correctly, which is less than the half. The results displayed that less

than half of the participants answered this item incorrectly, precisely 256 teachers. The ones answering this item incorrectly outnumbered the ones giving a correct answer. There are 49 teachers who did not know whether this item is true or false. The reason behind these findings in this item could be due to the teachers' sensitivity to errors. Most teachers cannot ignore the errors, especially non-native teachers are more sensitive to the errors. To put it differently, native teachers are more tolerant of student errors than non-native teachers of English (Sheorey, 1986; Schmitt, 1993; Rao & Li, 2017). Another reason could be the belief that the teachers are the figures who should be role models for their learners, and who provide right input for the learners. Because of these, the teachers may have thought that spelling errors should not be ignored. It is the teaching part; however, the logic behind the assessment part is different. Here, the assessment is on reading, and whether learners could understand the reading text or not. On the other hand, if errors of spelling are penalized, then the teacher is not assessing reading skills of learners, but rather something else. Furthermore, spelling also requires some kind of writing ability that should not be interfered with assessing reading.

Another item is "including not stated/doesn't say along with true/false items has advantages over true/false items". Heaton (1990) stated that true/false items are one of the most widely used tests for assessing reading skills. These tests are easy to prepare, and scoring is straightforward which could be some of the advantages of these tests. Despite having some advantages, true/false tests have also certain disadvantages one of which is 50% chance given to learners for guessing (Salend, 2009). As these type of tests support guessing of learners, it is a good idea to include another option to decrease the chance level of learners, and some of the alternatives for the third option could be *not given*, *not stated* or *don't know* (Hughes, 1989; Alderson, 2000). 43% of the respondents' answers are correct, less than the half. 236 participants' answer is correct whereas 221 teachers' answer is incorrect. In addition to these, don't know was selected by 85 teachers who were not sure if the answer was true or false.

"Simplification of reading texts is avoided" is one of the items that has the lowest mean score whose answer is false. Alderson (2000) and Hubley (2012) expressed that the texts should be checked in terms of their length or readability, then if necessary, some modifications should be done to adapt the reading text according to the language levels of the learners. In parallel with the previous statements, Heaton (1990) drew attention to the fact that learners have to be exposed to the materials in which the target language is

used for real purposes in real-life situations. However, learners may have difficulty understanding these authentic materials; hence, the thing that should be done here is to simplify the language of the reading text in terms of syntax, vocabulary and modify it based on the language level of the learners. This item was answered correctly by 37% of the teachers (N=205). The ones answering this item incorrectly (N=243) are more than the ones answering correctly. The number of the teachers who chose don't know option is one of the highest with 94 teachers. To put it differently, the teachers whose answer is incorrect outnumbered the ones who answered it correctly. The reason could be that the teachers might not be familiar with the modifications of the texts, or in their institutions, they might be using ready-made materials rather than simplifying. Due to the possibility that simplification is not a common practice and not an easy task, the teachers may have regarded it something that should be avoided.

Another item having one of the lowest mean scores is "the more items a reading text is followed, the more reliable it becomes" whose answer is true. Osterlind (1989), Heaton (1990) and Green (2014) stated that the reading texts that are short in terms of items are less reliable when compared to the texts having more items. This item was answered correctly by 36% of the respondents. When the answers are analyzed in detail, it was seen that out of 542 teachers, only 198 of them answered this item correctly. The rest 344 teachers were composed of the ones answering incorrectly (N=200) and selecting don't know option (N=144). The teachers choosing don't know option is one of the highest in assessing reading. It means that most of the participants either know that item incorrectly or do not know that one of the ways of making a reading text more reliable is increasing the number of questions following it.

Another low mean score belongs to the item "in a reading exam, using a text learners have encountered before is not a problem". The answer to this item is false. Hughes (1989) and Hubley (2012) stated that the texts used before should not be used again in a reading exam. Instead, the reading texts which are used in a reading exam could have familiar topics with the ones covered in the class or encountered in previous exams (Harris, 1969). Most of the participant teachers' answer was incorrect to this item, and these teachers who answered correctly make up 35% of the total number (N=190). There are 278 teachers answering this item incorrectly, which is more than the ones answering correctly. 74 participants did not know whether this item is true or false.

"When asking several questions about a reading text, all the questions are independent of each other" has the lowest mean score out of 15 items in assessing reading. The answer to this item is true. Osterlind (1989) and Alderson (2000) stressed the importance of designing independent questions related to a reading text. If the questions are not independent of each other, and they are somewhat related, then there is the risk that the answer of one question may affect the answer of the other question. In other words, one item may influence or determine the answer of the other item, because of this, independent questions related to reading texts should be prepared. Salend (2009) called this as similarity cues, and stated that because of these similarity cues, one information in one question leads to the answer in other question. Thus, they should be avoided. 28% of the learners (N=153) answered this item correctly, and the rest, 78%, either answered this item incorrectly (N=343) or did not know the answer of this item (N=46). So, it can be concluded that the participant teachers are not knowledgeable enough related to the features of well-prepared questions.

5.2.1.2. Assessing listening

The highest mean score belongs to the item "in selective listening, learners are expected to look for certain information". Selective listening is defined as scanning certain information (Rost, 1990; Brown, 2003; West & Turner, 2009; Rost, 2011). In selective listening, searching the overall or global meaning is not important, and apart from that specific information, learners do not need to understand the other parts of the listening text. Some tasks in selective listening may include listening to names, numbers, or certain facts. As is clear, learners face with a limited quantity of input regarding listening in selective listening. Buck (2001) also stated that the tasks in this group do not require learners to process the meaning; rather, learners need to understand specific information in the listening text. 58% of the participant teachers answered this item correctly. In total, there are 315 teachers answering this correctly whereas 187 participants gave an incorrect answer. It is the item in which the lowest number of the teachers selected don't know option. 40 participants were aware of the fact that they did not know the answer of this item. It is the item whose mean score is the highest; however, it is still an item which slightly more than half of the teachers answered correctly. It can be said that more than half of the respondents know the term selective listening which is one of the widely used techniques in assessing listening skills.

"A listening cloze test is a way of selective listening" is the item that has the second highest mean score. Brown (2003) stated that in a listening cloze test, learners are given a transcript in which there are certain gaps, and they are expected to fill in these gaps by listening to the words/phrases they hear. These listening cloze tests could be a story, monologue, or conversation. As is seen, this item is related to the first item, that is, they are both related to selective listening; hence, it can be concluded that the participant teachers have adequate knowledge regarding selective listening. Buck (2001) and Rost (2011) also uttered that as cloze tests are very popular, it is not very surprising to use them in assessing listening in which learners are required to fill in certain words or phrases which are deleted in the text. 52% of the participants' answer is correct for this item. 286 teachers knew the answer; however, 139 participants knew the answer incorrectly. In this item, there are 117 teachers who chose don't know option, which is really high in number. Why the number of the teachers choosing don't know option is high could be because this item includes terms such as *cloze test* and *selective listening*. The participants may not know what they mean. The previous item is also related to selective listening, and the number of the teachers choosing don't know is 40. Thus, the high number of don't know may not result from the term selective listening, but result from cloze test. The other possibility is that the teachers may know what should be done in selective listening, but they may not know the tasks included in selective listening. As a conclusion, when the number of the teachers answering this item correctly is investigated, it can be said that more than half of the participant teachers know what selective listening is, and are aware of the tasks included in selective listening.

The item which has the third highest mean score is "giving learners a transcript of the listening text is a valid way of assessing listening skills" whose answer is false. Buck (2001) gave the following example: Learners are given the transcript of a song, and after listening to a song, they are expected to fill in the gaps in the transcript. Indeed, recognizing words in the song is the purpose in this activity. However, the risk is that many of the gaps could be filled even without listening to the song. It is not a listening activity any more, but a reading activity in which learners have more chances to fill in the gaps by guessing or inferring meanings from the context. What Brown (2003) said was in parallel with the previous sentences. Brown (2003) drew attention to the weakness of this kind of activities in which the transcript of the listening text is given to learners by saying that this listening text is for sure a reading comprehension task from now on. As

Heaton (1990) mentioned, learners may not even listen to the text, or could understand a little but still they could fill in the gaps with the appropriate words based on their reading comprehension or general language ability. 47% of the participants (N=259) answered this item correctly, which is less than half of the total number. The number of the participants answering incorrectly (N=224) is very close to the ones answering correctly. The number of the teachers selecting don't know option is 59. It is obvious that nearly half of them know that they should not give learners the transcript of the listening text. The reason might be that there may be some listening exercises in which the transcript is given to learners in the coursebooks. As the teachers come across such examples, they may have thought this is the right way of assessing listening skills. However, the success rate is still low though it has the third highest mean score in assessing listening.

One of the items having the highest mean score is "including redundancy (e.g. what I mean to say is that ….) in a listening text poses a problem". The answer to this item is false because using redundancy in a listening text is not a problem. One of the features of listening input is that the input should include spoken language, and the language should be natural (Madsen, 1983; Rost, 1990; Aryadoust, 2013). As redundancy is a characteristic of everyday real language, and as the desired goal is to make the listening text authentic and natural, then redundancy should be a part of the listening texts used while assessing listening. In addition to these, redundancy can be used to replace any missing information in spoken language, and with the help of redundancy such as "what I mean to say…", the speaker has the chance to replace the missing information by using the language itself (Hughes, 1989; Heaton, 1990; Buck, 2001; Brown, 2003). 42% of the respondents answered this item correctly which might be regarded as low though the mean score is one of the highest. The exact number of people giving a correct answer is 228; on the other hand, there are 243 teachers giving an incorrect answer. 71 respondents stated that they did not know the answer of this item. It is obvious that the teachers answering incorrectly outnumbered the ones answering correctly. This finding shows that less than half of the participants know that the text to be used in a listening exam should include the features of spoken language including redundancy.

The next item is "any type of listening text is used for note-taking" the answer to which is false. Brown (2003) expressed that note-taking is a skill mostly used in academic world and classroom lectures; so, the texts to be used in note-taking should have certain features. Hughes (1989) and Heaton (1990) also stated that in note-taking, learners are

required to write down their notes while listening, and then they are given the questions related to the text used for note-taking. Because of this reason, the texts to be used for note-taking should be suitable texts from which notes could be taken successfully. Additionally, they stated that the short passages are not ideal to be used in note-taking; rather, longer passages should be selected in assessing listening. Madsen (1983) also expressed that the text to be used in note-taking should be selected carefully, and it should not be based on the general knowledge of learners. This item was answered correctly by 41% of the participants (N=223). 267 teachers answered this item incorrectly, and the number of the participants choosing don't know option is 52, which is one of the lowest in assessing listening. More than half of the respondents thought that any type of listening text could be used in note-taking. The reason behind this might be the misconception and the possible wrong practice of the teachers, because there might be teachers who tend to use each and every text for note-taking without being aware of the unique features of note-taking.

Having the sixth highest mean score, the next item is "in discrete-point testing, comprehension is at the literal/local level". Buck (2001) expressed that in discrete-point testing, we have the chance to divide and separate parts from each other, and each of these parts could be tested separately. Because of this separation and isolation, comprehension "is seen as understanding language on a local, literal level" (Buck, 2001, p. 66). Douglas (2010) and Flowerdew and Miller (2012) also uttered that as the emphasis is on the recognition of isolated elements, these isolated parts are tested independently, and, what is more, these are all treated as separate entities; thus, comprehension is at the local level. 36% of the teachers answered this item correctly. The exact number of them is 199, and the teachers answering incorrectly is 45. The results obtained from this item showed that the number of the teachers selecting don't know option is the highest in assessing listening (N=298). In other words, the number of the participants choosing don't know option outnumbered the ones answering correctly and incorrectly. It could be due to the terminology used in this item, which is discrete-point testing. It is possible that most of the participants may not know what discrete-point testing is.

"Using dictation diagnostically in assessing listening skills does not pose a problem" is the next item whose answer is true. Dictation is a preferred way of assessing listening skills of learners (Buck, 2001). As dictation is a type of integrative test of listening, they have been used a lot by teachers for many years (Brown, 2003). Through

dictation, teachers could have an idea about the performances of the learners in different areas such as spelling, grammar, writing, and cohesive elements (Buck, 2001; Brown, 2003; Flowerdew & Miller, 2012). With the help of dictation, teachers have the opportunity to detect the areas in which learners have weaknesses, and then could deal with these weaknesses. 31% of the respondents (N=171) were right in this item whereas 172 of them gave an incorrect answer. There existed 199 teachers who did not know the answer of this item. The probable reason is that the teachers may not know what dictation means or why it is used. It can be concluded that though dictation is one of the most widely used techniques in assessing listening, many of the participants do not know the reason why and when dictation is utilized.

"Errors of grammar or spelling are penalized while scoring" is the next item whose answer is false. Hughes (1989) stated that our main purpose is to assess listening skills of the learners, and we are sure that these learners have heard the correct word or phrase, but cannot write it correctly in terms of grammar and spelling; hence, we should ignore these grammar and spelling mistakes. Madsen (1983) also discussed that the purpose is to test learners' understanding of a piece of information or a text. Because of this, the vocabulary and grammar learners know should be used in listening tests, and learners' errors related to them should not be penalized (Cash & Schumm, 2006). Also, teachers assess grammar knowledge of the learners in a separate heading in the exam. If their grammar or spelling errors are penalized, then what they do is not assessing listening, but assessing grammar once more under the heading of assessing listening. 31% of the teachers said that errors of spelling and grammar should not be penalized, and they were right (N=169). However, the ones who just said the opposite (N=319) outnumbered the ones who answered correctly. The ones stating they did not know the answer of this item are 54 teachers. It might be because of the notion that the teachers may tend to cut scores from the grammar or spelling errors because of their being role models for their learners who provide right input. Another possible reson could be the teachers' sensitivity to errors, and they may not ignore the errors.

"Using reading texts for listening purposes poses a problem" is another item in assessing listening part of the scale. The answer to this item is true, because reading and listening are separate skills and their characteristic features are also different from each other. As Chafe (1985, cited in Buck, 2001) stated there exist certain differences between spoken and written language. In spoken language, the syntax is not as difficult as the

written form, and people tend to make use of shorter idea units whereas written language tends to include more dependent and subordinate clauses to convey more information. In spoken language, there are conjunctions as well, but they are usually simpler conjunctions such as *but*, *and*. However, in written language, more complex conjunctions are highly preferred. The spoken form also includes personal uses of language such as *I think*, *I mean*, and in spoken form, speakers have the tendency to show and share their feelings by making a direct reference to the listener. Thompson (1995) also expressed that when preparing listening texts there are some factors to take into consideration one of which is that the listening text should be close to oral rather than written form. Additionally, Heaton (1990) discussed that real life speech has certain features such as spontaneity, redundancy, hesitations, false starts and sometimes ungrammatical forms, and these features are missing in the written texts which are planned to be used as written texts to be read aloud. Madsen (1983), Mead and Rubin (1985) and Hughes (1989) also stated that the texts originally intended for reading should be avoided in assessing listening. 29% of the participants (N=160) stated that reading texts should not be used in assessing listening. It is clear that 71% of the teachers are not aware of the difference between the features of the spoken and written form of language. They either knew it incorrectly (N=292), or did not know the answer (N=90). It is seen that the number of the teachers giving an incorrect answer is much more than the others for this item.

"Scoring in note-taking is straightforward" is one of the items whose mean score is relatively low. Scoring in note-taking is really difficult indeed, and not an easy thing. As note-taking has many parts, it is challenging to score the note-taking tasks of learners. Brown (2003) attracted attention to this phenomenon by saying that the scoring process in note-taking is time-consuming owing to the reason that it is really subjective, and added that as the scoring system is subjective, reliability is lacking in scoring. Madsen (1983), Flowerdew and Miller (2012) and Green (2014) also stated that the scoring is really subjective in note-taking because there are no agreed-upon pre-determined answers as in objective tests, because of this reason, its reliability is less than objective tests. 24% of the teachers thought that scoring is not straightforward, rather a complex issue in note-taking. The exact number of the respondents answering correctly is 132 whereas there are 253 incorrect answers. This item has one of the highest numbers in don't know (N=157). This high number might result from the insufficient knowledge related to note-taking. Note-taking may not be a task used often in listening exams, or they may have insufficient

knowledge related to scoring, because note-taking might be an activity used in in-class assessment, not for scoring purposes. It can be concluded that teachers are not knowledgeable enough in terms of scoring in note-taking.

"Asking learners to listen to names or numbers is called intensive listening" is the next item whose answer is false. When specific information is the focus, then it is called selective listening (Rost, 1990; Brown, 2003; West & Turner, 2009; Rost, 2011). However, intensive listening is listening for the components of a larger unit such as words, discourse markers, intonation. In addition to this, the information at the recognition level is sought in this type. Morley (1972) and Buck (2001) discussed that the tasks are really helpful especially when learners have a problem with a specific sound or word. 23% of the participants knew what intensive listening is (N=126). Among the rest of the teachers, 278 teachers gave an incorrect answer to this item; however, there were 138 participants selecting don't know option. Hence, it can be said there are 138 teachers who are aware of the fact that they do not know the answer of this item. This high number in don't know may be due to the fact that the participants may not know what intensive listening is. When the items in assessing listening are analyzed, it is seen that the items having the highest mean scores are related to selective listening. Thus, it can be concluded that the respondents may have more knowledge in selective listening than intensive listening.

"Inference questions based on intelligence are avoided in listening tests" is the next item. Buck (2001) mentioned passage dependency in his book, and stated that there is passage dependency if the task could be completed after fully understanding the text. However, when learners rely on their background knowledge or intelligence, and based on these they could complete the task, then it is obvious that there is not passage dependency here (Mead & Rubin, 1985). The important thing here is that questions related to the listening text should not be predictable (Madsen, 1983; Hughes, 1989; Brown, 2003). Buck (2001) said that "we should include anything that is dependent on linguistic knowledge, and we should attempt to exclude anything that is dependent on general cognitive abilities" (p. 113). Moreover, Douglas (2010) stated that if such questions are included in an exam, then the reliability of this exam decreases. Because of this, questions based on intelligence is not the desired thing here; thus, they should be avoided. 18% of the teachers answered this item correctly (N=100) whereas 399 of them gave an incorrect answer. The number of the people choosing don't know option is 43,

which is the lowest in assessing listening. It was demonstrated that most of the participants do not have enough knowledge related to the idea that the primary concern is assessing listening, not something else such as intelligence or general knowledge, because these things have nothing to do with listening skills.

Another item is "spelling errors are ignored in scoring the dictation". As it is for sure that dictation is not a sole spelling test, the primary concern is not on spelling, and dictation is not developed as a form of spelling test (Madsen, 1983; Buck, 2001). Owing to this reason, spelling errors should be ignored in scoring dictation, because the learner has heard the word correctly, but has problems while writing it. As the primary concern is on listening, teachers should not cut the score of the learners due to spelling mistakes. Heaton (1990) and Brown (2003) also stated that dictation is highly preferred in assessing listening in which learners' both listening and writing skills are integrated. Due to this, as the focus is on assessing listening skills of the learners, whether the learners have heard the words or phrases correctly is of primary concern rather than being able to write each and every letter in these words and phrases. 16% of the teachers (N=92) gave a correct answer to this item. 400 teachers' answer is incorrect, and 50 teachers selected don't know option. It could be said that the teachers may not know that spelling errors should be ignored in scoring dictation. The reason behind this could be that though dictation requires many skills for the learner, teachers may still use them as a spelling test. Again, because of the possible wrong practice, they may have focused on spelling errors of the learners. What can be concluded here is that there are three items related to dictation in assessing listening. This item got the highest incorrect answer. The other two items had some of the highest answers in don't know. Thus, it can be said that the teachers may have difficulty in dictation and why it is used, and they have limited knowledge regarding dictation.

"Phonemic discrimination tasks (e.g. minimal pairs such as sheep-ship) are examples of integrative testing" is an item having one of the lowest mean scores. Phonemic discrimination tasks are one of the widely used tasks for testing listening in discrete-point approach and, in phonemic discrimination tasks, learners listen to one word in isolation and are expected to identify the word they have heard (Buck, 2001). These words are usually in the form of minimal pairs in which only one letter is different. As these are testing the parts of language, they cannot be accepted as a form of integrative testing, in which learners are required to make use of many elements at a time (Oller,

1979). Rather, these are a form of discrete-point testing in which elements of language are assessed independently. Madsen (1983), Hughes (1989) and Douglas (2010) called this type of task as the one in which the lowest abilities are assessed, and added that as the purpose is to distinguish the letters, they cannot be an example of integrative testing. Hughes (1989) also suggested that these tasks could be used for diagnostic purposes. 11% of the participants answered this item correctly (N=63) which is the second lowest mean score. There are 209 teachers answering this item incorrectly. The number of the teachers choosing don't know option is 270, which is one of the highest in assessing listening. The teachers may have difficulty in understanding integrative testing. The terms that are phonemic discrimination task and integrative testing may have led to this high number in don't know. Since an example is given for phonemic discrimination task, it may not be a problem for teachers' understanding. Because even if they may not know the term, it is possible that they are familiar with the task regarding phonemic discrimination tasks in coursebooks. Based on this possibility, the teachers may have had problems related to integrative testing.

"Dictation is a kind of discrete-point testing" is the last item which has the lowest mean score out of 15 items related to assessing listening. As Buck (2001) stated, dictation is "the most widely used integrative test of listening" (p. 73), and added dictation is not only a listening test, it is more than a listening test. Short-term memory is involved in this process, and writing ability is required for learners to be able to write what they have heard; hence, it is clear that it is more than a listening test. Heaton (1990) also expressed that dictation is an integrative test because it "measures a complex range of integrated skills and should not be regarded simply a test of spelling" (p. 151). It is integrative because in integrated way of assessing, there is a shift from the discrete measurements of language items, and more than one item are tested at the same time (Madsen, 1983; Douglas, 2010; Flowerdew & Miller, 2012). Brown (2003) also said that some sophistication is needed in dictation, and learners are expected to have some grammar knowledge and discourse expectancies, and because of all the reasons mentioned above, dictation is not a kind of discrete point testing; rather, it is a kind of integrative testing. 52 of the respondents gave a correct answer to this item whereas 253 of them answered incorrectly. The number of the participants answering incorrectly outnumbered the ones answering correctly. Besides, there are 237 participants choosing don't know option, which is again too high. As mentioned above, it was found that the participant teachers

have limited knowledge regarding dictation, and they also have limited knowledge related to test types such as discrete-point, integrative, etc. In practice, teachers may tend to use dictation to check whether learners are able to understand and write words and phrases or not. Maybe because of this possible common misuse of dictation, the teachers may have thought that dictation is simply a spelling test, or maybe they may not know the details related to discrete-point testing and what tasks are included in this. Thus, the problem may stem from the terminology used in this item. As a result, it is clear that teachers have difficulty in the items related to integrative and discrete-point testing.

5.2.1.3. Assessing writing

The highest mean score belongs to the item "using visuals which guide learners for writing poses a problem". Using pictures or visuals are a great tool for learners to write about (Ruth & Murphy, 1988; Hughes, 1989; Heaton, 1990; Douglas, 2010; Weigle, 2012). Heaton (1990) believes that using visuals has two main advantages. One is stimulating learners' imagination via visuals, and the second one is that learners tend to use the sentences included in the verbal stimulus; however, with the help of visuals, this weakness diminishes. 77% of the teachers (N=422) gave a correct answer to this item, which is a high number. There are 50 teachers answering this item incorrectly, and 70 teachers chose don't know option. It is clear that most of the teachers know that visuals could be used as prompts for a writing task.

"The parts of a scoring scale and the scores in each part do not change for different levels of learners" is the item that has the second highest mean score. An existing scale should not be used for all learners and levels, and it should be modified based on the purposes and needs of the learners. Harris (1969), Hughes (1989) and Heaton (1990) stated that the parts of a scoring scale and the weighting given to each part should be modified, and necessary changes should be made at various levels. In parallel with the curricular goals and learners' needs, the parts of the scoring scale could be tailored, and with the help of these modifications, some parts may be given more emphasis at different levels (Bachman & Palmer, 1996; Weigle, 2002; Brown, 2003; Fulcher, 2003). In other words, based on the needs of the learners, the parts and the scores in each part should be modified at different levels because the needs of the learners are all different at different levels. 61% of the respondents answered this item correctly (N=335). 150 teachers gave an incorrect answer to this item, and 57 of them did not know the answer. It shows more

than half of the participants know that some modification is needed before using scales based on the purposes and the language level of the learners.

"Giving restrictive prompts/guidelines to learners for the writing task is avoided" got the third highest mean score of 15 items in assessing writing. The answer to this item is false. Bachman and Palmer (1996) came up with three guidelines for instructions of writing tasks, and one of them is providing clear, restrictive and detailed instructions in which there is information related to specification of audience, purpose of writing, and how long the response will be. Douglas (2010) stated that giving clear instructions to learners for a writing task is crucial, because thanks to restrictive prompts/guidelines, variations in scoring decrease (Weigle, 2002). This specification or restriction is needed (Ruth & Murphy, 1988; Salend, 2009; Green, 2014) because it increases reliability in scoring, since what is expected from learners is written in detail (Hughes, 1989). Harris (1969) and Heaton (1990) also mentioned that scoring becomes more reliable because it provides an opportunity for teachers to compare different compositions more easily across learners, and added that these tasks with restrictive prompts/guidelines provide a washback effect on teaching and learning while preparing exams. 61% of the teachers gave a correct answer to this item, the same mean score with the previous item. The exact number of the respondents giving a correct answer is 333 whereas 155 participants answered this incorrectly. 54 teachers stated that they did not know the answer, which is one of lowest numbers in assessing writing. This result indicates that more than half of the participant teachers know that learners should not be set free, rather, they should be limited and restricted by means of instructions or guidelines. With the help of these restrictions, the learners could understand what is expected from them better, and teachers could also have a chance to compare the written works across learners.

"Irrelevant ideas are ignored in the assessment of initial stages of a written work in process writing" is the item that got one of the highest mean scores in assessing writing though it is still low in mean score in general. The answer to this item is false. Brown (2003) came up with some guidelines for assessing the initial stages of writing, and one of them is focusing on the meaning and main points, and added that grammatical and lexical errors should be dealt with at later stages, not at initial stages. In other words, the focus should be on the identification of irrelevant ideas, and whether the sentences are relevant or not, and the primary goal is what learners can write and what cannot write (Heaton, 1990; Weigle, 2002; Bright, 2007; Chapman & King, 2009). 53% of the

participants (N=292) know the answer of this item, which is slightly more than the half. There are 173 teachers giving an incorrect answer, and 77 teachers selecting don't know option. It is clear that slightly more than half of the participants have the knowledge that the focus is on whether the ideas of the learners are relevant or irrelevant in their written works at the initial stages, and the relevancy of the ideas should be dealt with from the very beginning.

"Analytic scoring is used to see the strengths and weaknesses of learners" is the next item the answer to which is true. Analytic scoring should be used if the purpose is to seek diagnostic information about the learners (Hughes, 1989; Heaton, 1990; Bachman & Palmer, 1996; Brown, 2003). This diagnostic information tells a lot in relation to learners' strengths and weaknesses, and a learner may be good at certain aspects but has flaws in others; hence, analytic scoring gives the teacher the possibility to detect the strengths and weaknesses of learners (Huot, 1996; Weigle, 2002). 51% of the participants answered this item correctly, slightly more than the half. The number of the teachers giving a correct answer is 279 whereas 177 teachers answered it incorrectly. 86 participants selected don't know option. It can be concluded that half of the teachers in the current study are aware of what analytic scoring is and why it is utilized. Most of the teachers are required to use a type of a scoring scale in most of the institutions, and this choice may depend on the purpose. However, what can be drawn from this finding, only half of the participants are conscious about the purposes of this scoring scale.

The next item is "in controlled writing, learners have the chance to convey new information" whose answer is false. Brown (2003) stated that certain assessment tasks have a concern for form, and they are strictly restricted and controlled by teachers or test designers. Some of the tasks included in this group are picture-cued tasks, ordering tasks, short answer tasks, etc. As they are strictly controlled, it is not possible for learners to convey new information through these tasks. Controlled writing is also called as guided writing in which there is no new information transmit (Madsen, 1983; Heaton, 1990; Silva, 1990; Brown, 2003, Fulcher, 2003). This item was answered correctly by 48% of the participants (N=261), which is slightly lower than the half. 163 teachers answered this item incorrectly. There are 118 teachers who chose don't know option. This number is one of the highest ones in don't know in assessing writing. Why a lot of people do not have any knowledge related to this item could be the use of controlled writing. Maybe the teachers do not have enough knowledge about the term controlled writing; thus, they

chose don't know option. Generally, the results in this item mean nearly half of the teachers know what controlled writing is.

Another item is "holistic scoring is used to see whether the learner is proficient or not at the end of the term" whose answer is true. At the end of the term, the purpose is not to get diagnostic information through tests; rather, the purpose is to see whether learners have passed the criterion level or not. The other name of holistic scoring is impressionistic scoring in which the overall impression is the basis of the scoring which is used in determining the proficiency levels of the learners, because the concern is not whether learners are good or bad at certain parts of the scoring, but rather, whether they get the satisfactory score or not is of concern (Hughes, 1989; Weigle, 2002; Brown, 2003; Shaw & Weir, 2007; Weigle, 2012). 47% of the respondents gave a correct answer to this item. 257 teachers answered it correctly; on the other hand, there are 161 participants giving an incorrect answer. The number of the people choosing don't know option is high in number (N=124). Though this mean score of the ones giving a correct answer is one of the highest ones, it is still a low one. This finding displays that only half of the participant teachers know the purpose why and when holistic scoring is used.

"Providing a reading text for writing is a way of assessing writing skills" is another item, and the answer to it is true. A reading text could be provided to learners, and based on this, they are asked to write their tasks (Ruth & Murphy, 1988; Hughes, 1989; Heaton, 1990; Weigle, 2002). Weigle (2002) stated that while giving a reading text for all learners, a common basis is created for the learners, and they are all equal in terms of relying on the same text and being given the same input. Another point is that learners do not have to think what they will write about and produce ideas, and as a starting point, they are given the reading text, and furthermore, the reading text could activate their background knowledge, and it will be much easier for them to come up with ideas (Weigle, 2002). 46% of the respondents gave a correct answer to this item, less than the half. The number of the participants answering correctly is 250 whereas there are 196 teachers giving an incorrect answer. 96 of the teachers selected don't know option, and these teachers are aware of their missing knowledge related to this item. Thus, it can be concluded that more than half of the teachers do not know that a reading text could be used as a prompt for a writing task.

The next one is "analytic scoring leads to greater reliability than holistic scoring in writing" which is true. In analytic scoring, the rater has to give separate scores for various

parts of the scale which in turn makes the scoring more reliable (Hughes, 1989; Heaton, 1990; O'Sullivan, 2012; Green, 2014). Bachman and Palmer (1996) came up with a framework of test usefulness in which it is stated that reliability in analytic scoring is higher than holistic scoring. This item was answered correctly by 39% of the participants (N=216). It was answered incorrectly by 192 of them, and 134 teachers selected don't know option. The number in don't know is one of the highest in assessing writing. The reason for this tendency could be the terms used in this item that are analytic scoring and reliability.

"Classroom evaluation of learning in terms of writing is best served through analytic scoring rather than holistic scoring" is another item the answer to which is true. Brown (2003) stated that the choice of rating scales is based on the purpose, and with analytic scoring, teachers get more detailed information related to learners' performances. He also added that when analytic scoring (in which many elements are scored) is used in classroom evaluation, teachers have more chances to detect the weaknesses and strengths of learners. Teachers also get the opportunity to tailor their teaching based on the feedback they obtain from the performances of learners on the writing tasks through analytic scoring (Weigle, 2002). Moreover, as analytic scoring is more informative than holistic scoring, for classroom purposes, analytic scoring should be used (Heaton, 1990; Shaw & Weir, 2007; Weigle, 2012; O'Sullivan, 2012). 39% of the respondents (N=214) know that for classroom evaluation it is better and more appropriate to use analytic scale rather than holistic scale. 167 of the participants gave an incorrect answer to this item, and 161 of the teachers selected don't know option. The number in don't know is one of the highest in assessing writing. When the high numbers in don't know are investigated, it is seen that these items are either related to analytic or holistic scoring. The teachers tended to choose don't know option more in the items concerning analytic and holistic scales compared to the other items in assessing writing. It is obvious that the teachers have difficulty in the types of scales used in assessing writing. As a conclusion, less than half of the teachers do not have enough knowledge related to the use of analytic or holistic scoring.

"Mechanical errors (e.g. spelling and punctuation) are dealt with in the assessment of later stages of a written work" is the next item. Brown (2003) suggested that after the teacher deals with irrelevant ideas at initial stages, at later stages, fine-tuning is necessary in which grammatical and mechanical errors are dealt with. Heaton (1990) also stated that spelling and punctuation should not be of primary concern, and added that the most

important thing is that whether learners could express themselves or not by using relevant ideas and appropriate kind of language. Additionally, Weigle (2002) and Chapman and King (2003) stated that learners are encouraged not to focus on mechanical errors in the initial stages, because focusing on mechanical errors hamper their flow of ideas. This item was answered correctly by 34% of the teachers (N=189). The rest either gave an incorrect answer (N=298), or selected don't know option (N=55). The number of the participants answering this item incorrectly outnumbered the ones answering correctly. The results demonstrated that even though the teachers are more knowledgeable in dealing with irrelevant ideas in the initial stages of a written work, they are not knowledgeable enough in dealing with mechanical errors at later stages. The reason behind this could be that, as stated earlier, the teachers may have thought that they should deal with all kinds of errors from the very initial stages of a written work due to the sensitivity to errors. As is stated in the literature (Sheorey, 1986; Schmitt, 1993; Rao & Li, 2017), non-native teachers are not very tolerant of errors of learners. Another reason could be that the teachers may be used to evaluating a written work from different angles based on the parts in the scoring scale. Relying on this practice, they may have thought that mechanical errors should be dealt with from the very beginning.

"Giving two options to learners and asking them to write about one ensure reliable and valid scoring" is an item having one of the lowest mean scores. Indeed, options should not be given to learners. Weigle (2002) and Heaton (1990) expressed that when learners are given two options, they waste time on which topic they are going to write. This time could be spent in writing, but in this case, learners spend this time with choosing one option. Weigle (2002) also stated that the stronger argument for this is providing tasks that are equal in difficulty is a big problem, and how to measure difficulty is another big problem. In another book, Weigle (2012) mentioned another problem related to giving options to learners. Providing options to learners makes scoring less reliable, because different learners will write on different topics resulting in the difficulty comparing across the writings of learners. In other words, if no choice is given to learners, then comparisons between learners will become easier (Oller, 1981, cited in Ruth & Murphy, 1988; Hughes, 1989). 29% of the teachers answered this item correctly, and the exact number is 312 whereas this number is 160 for the ones answering incorrectly. Apart from these groups, there are 70 teachers choosing don't know option.

"Learners are required to write about at least two tasks in the exam rather than one task" is an item having one of the lowest mean scores. Hughes (1989) stated that learners have to be given as many tasks as possible. In fact, a valid writing exam should include the tasks in which learners have the chance to perform all the relevant tasks they have covered. However, it is a problem in terms of practicality; instead, learners should be required to write at least two tasks. By giving them chances through at least two tasks, learners are given as many fresh starts as possible (Oller, 1981, cited in Ruth & Murphy, 1988; Hughes, 1989). As a result, if learners are good at a specific task, but not as good as at others, by asking them to write at least two tasks this risk is eliminated. Heaton (1990) discussed that asking them to write at least two tasks yield more reliable results when compared to one task, and more than one register or genre could be tested at a time through at least two tasks. 27% of the participants (N=149) answered this item correctly, implying that 73% of them either gave an incorrect answer to this item or chose don't know option. 309 teachers answered incorrectly, and 84 participants selected don't know option. The reason of this could be that learners may be given one task rather than two in some institutions, thus leading the teachers to think that it may be the correct way of assessing writing skills. Because of this probable wrong practice, the teachers in the present study may have thought that one task should be given to learners.

Another item which got one of the lowest mean scores is "when there is a disagreement between the scores of the two raters, they score the written work again". Weigle (2002) stressed the importance of rater training by saying that rater training increases the reliability of raters. One of the important points raters should take into account is the scoring of the writing tasks. Raters, at least two raters, should score each writing script independently, and if there exists a huge difference between the scores of the two raters, then a third rater should read the script (White, 1984; Hughes, 1989; Weigle, 2002; Brown, 2003; Shaw & Weir, 2007; Weigle, 2012). After the scoring of the third rater, all three scores could be averaged or the score of the third rater will be the score given to the writing task (Weigle, 2002). Raters should not score their work again, because in this second turn, they will be affected by the discrepancies between their score, and their scoring will be influenced by them. Thus, the results will not be reliable any more. 24% of the participants gave a correct answer to this item. This mean score is the second lowest one. 381 teachers gave an incorrect answer to this item, and there are 27 participants choosing don't know option. This number of the teachers choosing don't

know option is the lowest in assessing writing. Thus, there is the tendency of the teachers for either true or false for this item, rather than don't know. It can be stated that most of the participant teachers do not have sufficient knowledge related to the need for the third rater when there are discrepancies between two raters.

The lowest mean score belongs to "giving learners an opinion and asking them to discuss it is a valid way of assessing their writing skills" whose answer is false. Harris (1969) and Hughes (1989) strongly suggested that teachers should assess learners' writing abilities, not their creativity, imagination, or the ability of making justifications for their opinions or not. Here, when learners are given an opinion and asked to write on this, teachers will be affected by the opinions of the learners. He added that teachers have a set of arguments in their minds while scoring the tasks, and none of the learners could meet these criteria of the teacher, and none of the learners could please the teacher fully, which in turn becomes a threat to valid assessment of writing skills. Heaton (1990) also suggested that scoring will not be very reliable, because there is not a limitation for the topic. As learners come up with any ideas, scoring will be a problem for the raters. 13% of the participant teachers (N=72) thought that giving an opinion and asking learners to write about this is not valid. 420 teachers answered this item incorrectly, and 50 participants selected don't know option. Hence, there is not a high inclination of the teachers for don't know in this item. It can be expressed that 50 teachers know that they do not have this knowledge in themselves related to this item, that is, these teachers are aware of their missing knowledge in themselves. Giving such a prompt to learners is not a valid way of assessing their writing skills, because it is natural for the teacher to be influenced by the opinions of the learners, which is a risk for the validity.

5.2.1.4. Assessing speaking

"Giving learners one task is enough to assess speaking skills" received the highest mean score of all the items. The answer to it is false, because one task is not enough to assess speaking skills of learners. It was stated in the literature that there should be more than one task in an exam representing a wide sample covered in the class (Madsen, 1983; Hughes, 1989; McNamara, 1996; Fulcher, 2003; Green, 2014). Hughes (1989) added that in these tasks, learners should be presented various formats at the same time, and given as many fresh starts as possible. 89% of the teachers (N=486) answered this item correctly. This mean score is the highest one among all items in the whole scale. 34

teachers gave an incorrect answer, and 22 of them selected don't know option. The number of the participants selecting don't know option is one of the lowest in assessing speaking. It is clear that nearly all of the teachers are aware of the fact that learners should be provided with at least two tasks in assessing speaking skills. On the other hand, these participants selected true in the item in assessing writing, which is "learners are required to write at least two tasks in the exam rather than one task". The logic behind both items is the same; but, the teachers gave different answers to these items. As expressed in the related item in assessing writing, the reason could be that in the institutions in Turkey, there might be a tendency to give only one task to learners in assessing writing. In assessing speaking, the institutions may not give at least two tasks to learners. Because of these possible wrong practices of the institutions, the teachers may have given these answers. This clash between these items may be a good indicator of learning from practice; thus, it might show how crucial the practices could be in shaping the knowledge of the teachers.

The second highest mean score belongs to "the interlocutor gives the score when the learner is in the exam room" the answer to which is false. When learners are trying to speak and accomplish the speaking tasks in an exam, teachers should avoid giving their score in the meantime (Harris, 1969; Hughes, 1989; Heaton, 1990; Luoma, 2004). 79% of the respondents gave a correct answer to this item. 430 teachers answered this item correctly whereas 72 participants answered it incorrectly. There are also 40 teachers selecting don't know option. It can be concluded that most of the participants know that after the learner has left the room, the scores should be given. The reason could be that teachers may get distracted and may not focus on the utterances of the learners while trying to give the scores. Besides, the learners may also be distracted because they may not feel relaxed, they may focus on the score of the teacher, not what they are going to say, and this scoring may lead them to more pressure and stress.

"Interlocutors' showing interest by verbal and non-verbal signals poses a problem" has one of the highest mean scores whose answer is false. The teacher should have a sympathetic attitude towards the learners in a speaking exam, and should be supportive (Madsen, 1983; Hughes, 1989; Heaton, 1990; Brown, 2003; Luoma, 2004). Heaton (1990) also discussed that the aim of the interlocutor should be holding a real conversation. Based on this, in daily life, speakers show interests to each other as a sign of listening to the other and giving importance to the other's utterances. When teachers

show interests instead of sitting still during the exam, which is unnatural, the interaction between the teacher and learner becomes authentic. This item was answered correctly by 71% of the teachers (N=386). 125 teachers gave an incorrect answer to this item, and 31 of the participants chose don't know option, which is one of the lowest in assessing speaking. In other words, most of the teachers know that speaking is a genuine interaction, and it is natural to show interests while listening to a person.

The item that is "discussion among learners is a way of assessing speaking skills" is one of the items having a high mean score among 15 items. In discussion, learners are required to interact with each other on a given subject, and it is a way of assessing speaking skills of learners in which the participants share the responsibilities of the speaking task (Madsen, 1983; Hughes, 1989; Brown, 2003; Fulcher, 2003; Luoma, 2004; O'Sullivan, 2012). Heaton (1990) expressed that tasks in discussion are meaningful and require active learner involvement. Furthermore, he added that discussion is not a mechanical task; rather, learners try to communicate with each other. 57% of the respondents answered this item correctly, more than half of the number. The number of the participants giving a correct answer is 312 whereas 213 teachers gave an incorrect answer. 17 teachers chose don't know option which is the lowest number in assessing speaking. It could be concluded that more than half of the participants know that discussion is one of the ways of assessing speaking skills of learners.

"A checklist is a means of scoring oral presentations in in-class assessment" is the next item. With the help of checklists, learners are provided with detailed and diagnostic information related to their performance in classroom. These checklists could be plus or minus, or yes or no, and are really descriptive in nature so that learners could see clearly which aspects get the plus or yes and which aspects get the minus or no (Heaton, 1990; Brown, 2003; Luoma, 2004, O'Sullivan, 2012; Green, 2014). 53% of the participants (N=288) gave a correct answer to this item, slightly more than the half. 183 participants gave an incorrect answer, and 71 teachers selected don't know option. Nearly half of the teachers are aware of the use of checklists as a means of evaluating oral presentations in the class.

"In a speaking exam, production and comprehension are assessed together" is another item. In a typical interaction, people talk to each other, and these people are both speakers and listeners because they share the responsibilities of spoken interaction to convey their messages and speaking is meaningful (Madsen, 1983; Hughes, 1989;

Heaton, 1990; Brown, 2003; Luoma, 2004; O'Sullivan, 2012; Green, 2014). Heaton (1990) discussed that successful communication and interaction rely on both the speaker and the listener. In an ideal speaking exam, as the speaking and listening are interrelated, production and comprehension should be together; thus, the tasks in a speaking exam should assess both of them. This item was answered correctly by 52% of the teachers (N=282), and incorrectly by 231 teachers. The ones who selected don't know option are low in number (N=29). This finding suggested that nearly half of the participants know that speaking is a real interaction that involves listening as well. As the real purpose is to have a genuine interaction with the learners according to the recent trends, both comprehension and production should be assessed together.

Another item is "when the focus is to assess discourse, role plays are used" whose mean score is not one of the highest or lowest ones. In role plays, learners assume the role of somebody else, such as a passenger on a train or a person going to a restaurant and ordering a meal, and the purpose in role plays is to see whether learners could handle with these situations or not (Hughes, 1989; Heaton, 1990; Brown, 2003; Fulcher, 2003; Luoma, 2004; O'Sullivan, 2012). They also stated that the new role and situation are important for the learners' using appropriate language, and as role plays stimulate reality, language becomes a means in that new role and situation. 50% of the respondents gave a correct answer to this item, which is half of the total number. The exact number of the teachers giving a correct answer is 270, and 166 teachers gave an incorrect answer. The ones who chose don't know option are very high number, which is 106. Why the number in don't know is really high in assessing speaking may be due to the word discourse, maybe the teachers do not know the meaning of discourse. Thus, the teachers may have problems related to terminology, and they do not have enough knowledge regarding this word. The other possible reason could be the use of role-plays in assessing speaking. If role-plays are not used very often in some institutions, or may not be preferred by the teachers, then the teachers may have had limited knowledge related to them. This result shows that half of the participants know that if the purpose is to assess discourse, then role-plays can be used for that purpose.

Another item is "using holistic and analytic scales at the same time poses a problem". Holistic scales and analytic scales are used for different purposes. One type of scale can be used as a complement to the other so that the weaknesses of one scale can be compensated by the other kind of the scale (Hughes, 1989; Luoma, 2004). It was

answered correctly by 42% of the teachers (N=231), less than half of the teachers. 149 teachers answered this item incorrectly, and there are 162 participants choosing don't know option. The tendency towards don't know may be because of the names of the scales that are analytic and holistic. As mentioned in assessing writing, the participants tended to choose don't know option for the items including analytic or holistic scales. This tendency towards don't know in all the items related to the types of the scales makes it clear that the participant teachers lack enough knowledge concerning analytic and holistic scales. All the results for this item suggest that more than half of the participants do not know that these two scales could be used together to complement each other.

"In interlocutor-learner interviews, the teacher has the chance to adapt the questions being asked" is an item whose answer is true. Interviews are the most used tasks in a speaking exam, in which the interlocutor asks all the questions. Learners are expected to answer these questions, and the interlocutor has the opportunity for adaptation (Madsen, 1983; Hughes, 1989; Brown, 2003; Fulcher, 2003; Luoma, 2004; Douglas, 2010; O'Sullivan, 2012). Heaton (1990) stated that interlocutor should become flexible so that there is the chance to adapt and direct the language to be used during the speaking exam. 38% of the teachers (N=209) answered this item correctly. 277 of them answered incorrectly, and 56 teachers selected don't know option. The teachers giving an incorrect answer outnumbered the ones giving a correct answer. The reason could be that in some of the institutions, the teachers may have a limited number of prompts to be asked to the learners, and may not be allowed to make any changes on these prompts as a part of their institutional policy. Because of this possible wrong practice, the teachers may have thought that adaptation of the questions should be avoided.

Another item is "when the interlocutor does not understand the learner, giving that feeling or saying it poses a problem" the answer to which is false. Hughes (1989), Heaton (1990) and Fulcher (2003) suggested that the interlocutor should contribute to the speaking task without interrupting and dominating the learner's speech too much. 35% of the participant teachers (N=191) answered this item correctly, implying that many of the teachers in the present study do not have this knowledge. 308 teachers gave an incorrect answer to this item, and 43 participants selected don't know option. The teacher should tell the learner that something is wrong with her/his message, then the learner could paraphrase her/his messages or give details to make her/himself clearer. When the teacher does not give the feeling or the message that s/he has not understood, then the learner

may be given a low score as a result. However, when shown reactions verbally or nonverbally, the learner may be given a chance to compensate the previous utterance or to express her/himself in a clearer and more organized way.

"When it becomes apparent that the learner cannot reach the criterion level, the task is ended" is the next item. Harris (1969), Hughes (1989) and Brown (2003) stated that if it is clear that the learner cannot get the criterion level, the interlocutor should end the task, because there is a criterion and learners should be above that level. However, if the learner does not have adequate knowledge to be above that level, it does not seem okay to ask the same questions again and again to that learner, which may make that process longer. Moreover, if the task is not ended, then the learner may become more anxious and have negative feelings while the interlocutor waits for the answer from the learner in silence or repeats the same question. 28% of the teachers answered it correctly; that is, most of the teachers' answers to it were incorrect. The number of the teachers giving a correct answer is 157; however, there are 320 participants answering incorrectly. Besides, 65 respondents chose don't know option. The teachers are advised to be sympathetic and supportive during the speaking exam, and based on this, they may have thought that they are not being supportive enough by ending the task. However, the idea behind this is that there is a criterion set before the exam, and the learners are assessed based on this criterion level. If it is clear that the learners cannot reach it, then the task is ended. If it is not ended, what is expected from the learner may not be realistic because that learner cannot achieve this task so that that learner may feel more anxious and even feel discouraged in this situation.

"In interactive tasks, more than two learners pose a problem" is an item receiving one of the lowest mean scores. If more than one learner are included in a speaking task, and they are engaged in performing a task, then it poses a problem for less outgoing learners, because they will be suppressed by more dominant learners (Hughes, 1989; Heaton, 1990; Brown, 2003; O'Sullivan, 2012). Heaton (1990) also stated that if the utterance of one learner contains some mistakes, this will affect the comprehensibility of the other learner. Even two learners have these risks in terms of dominance, characteristic features or language level, three or more learners increase these risks, and it should be avoided. This item was answered correctly by 27% of the respondents (N=149). To put it differently, most of the teachers in this study thought that more than two learners in a speaking exam is not a problem. 316 teachers gave an incorrect answer to this item, and

there are 77 respondents choosing don't know option. The number of the people giving an incorrect answer to this item outnumbered the number of the teachers giving a correct answer.

"Asking learners to repeat a word, phrase or a sentence is a way of assessing speaking skills" is one of the items having a low mean score. The answer is true. One of the tasks to be used in a speaking exam is asking learners to repeat the words or sentences they hear (Harris, 1969; Madsen, 1983; Hughes, 1989; Heaton, 1990; Fulcher, 2003; Brown, 2003). Hughes (1989) stated that repetition could be used for specific words and sentences in which learners make the same types of mistakes. 20% of the participants (N=112) gave a correct answer to this item, implying that 80% of them either answered incorrectly or selected don't know. There are 359 teachers answering incorrectly, and there are 71 teachers selecting don't know option. It is clear that the number of the participant teachers outnumbered the ones giving a correct answer. 71 respondents were conscious about the fact that they do not have this knowledge related to this item.

The item having the second lowest mean score is "in peer interaction, random matching is avoided". For peer interaction, it is stated in the literature that the learner's speech is affected by the other learner's speech, personality, and communication style, and the language level of the learners is also influential in peer interaction (Hughes, 1989; Heaton, 1990; Weir, 1993; Brown, 2003; Fulcher, 2003; Luoma, 2004; O'Sullivan, 2012). Whether both learners in peer interaction have the equal opportunity to speak is a big concern for all the scholars mentioned. In terms of personality, it becomes difficult for introvert learners to shine in this task, because of this, learners should not be matched randomly in peer interaction, and teachers should make sure that the two learners have similar personality features and language levels (Heaton, 1990; O'Sullivan, 2012). This item, having the second lowest mean score, was answered correctly by 18% of the teachers (N=100). There are 342 teachers answering incorrectly, and there are 100 teachers choosing don't know option. This finding indicated that most of the teachers do not have adequate knowledge related to the matching of the learners in assessing speaking. The reason behind this may rely on the practices of the institutions in which learners may be matched randomly due to practicality reasons. As there are a lot of learners in preparatory programmes, it may not be very practical and easy to match all the learners consciously based on their language levels or personalities. Based on this

probable wrong practice, the teachers may have thought that random matching is appropriate.

"Reading aloud is a technique used to assess speaking skills" got the lowest mean score of all items. Reading aloud is a kind of structured speaking tasks in which learners are highly controlled (Harris, 1969; Madsen, 1983; Hughes, 1989; Heaton, 1990; Brown, 2003; Luoma, 2004, O'Sullivan, 2012; Green, 2014). In other words, teachers know exactly what the learners are going to say in this type of speaking task, because of this they are controlled. Luoma (2004) stated that as learners cannot come up with any unpredictable and creative language use, and as the input and the output are the same for all learners, scoring is really straightforward and scoring becomes more reliable because of the chance of comparability across learners. Heaton (1990) suggested that learners should be asked to read aloud the scripts that they are likely to encounter in their daily lives while communicating and interacting. 16% of the participants (N=87) know that reading aloud is one of the techniques to be used in assessing speaking; however, 84% of them either gave an incorrect answer to this item or selected don't know option. 380 teachers answered incorrectly, and 75 teachers selected don't know option. It is obvious that most of the teachers do not have enough knowledge related to reading aloud. The reason could be that the teachers may have thought that reading aloud is a technique used to assess a limited part of speaking skills, and because of that they may have chosen the wrong answer. However, it is not very easy and practical to divide a skill into parts. Rather, speaking skill as a whole includes many subskills some of which are spelling, intonation, pronunciation, etc.

5.3. The Relationship among the Participants' Skill-based Assessment Knowledge

The third research question tried to find an answer whether there exists a relationship among the teachers' levels of skill-based LAK. The findings revealed that all the items are correlated with LAK, implying that these four skills are the components of LAK, and if the knowledge level of the teachers increases in these skills, then the knowledge level in LAK tends to increase as well. Furthermore, the skills have high or moderate positive correlations among themselves, indicating that if EFL teachers' assessment knowledge in one skill increases, their assessment knowledge in others tends to increase in high or moderate levels. These results suggested that language is a holistic phenomenon even if it is composed of various skills. The probable reason for this is that

all the skills, though different in nature, serve for the same purpose which is LAK, and the logic behind the assessment of all skills is similar. For instance, when a teacher's knowledge in designing tasks such as multiple choice or open-ended in reading increases, that teacher can transfer this knowledge into other skills, and makes use of that knowledge in others. Another example could be the use of at least two tasks in assessing each skill. When a teacher has learnt that at least two tasks are needed to assess writing skills more reliably, then the same information could be used in other skills as well. Consequently, this increase in knowledge in one skill affects the knowledge in other skills positively, and also results in increased knowledge in LAK.

5.4. Effects of Demographic Features on LAK Level of the Teachers

The fourth research question investigated whether LAK level changes according to certain demographic features that are years of experience, the BA programme being graduated, educational background, workplace, testing course in BA, attending trainings on testing and assessment, and being a testing office member.

Whether the teachers' language assessment knowledge changed was searched based on each variable, and the findings displayed that the LAK level of the participants did not change according to years of experience, educational background, the BA programme being graduated, workplace, testing course in BA, and attending trainings on testing. The only difference was between the teachers who were in the testing office and who were not a member of the testing office. It can be concluded from these findings that years of experience, the BA programme being graduated, educational background, workplace, testing course in BA, and attending trainings on testing and assessment do not have an effect on LAK level of the participants whereas being a testing office member has an influence on LAK level of the teachers.

To start with years of experience, in Tao (2014)'s study, it was revealed that there is not a relationship between years of experience and the actual LAK level of teachers. Thus, this finding is in parallel with the results of the current study. The possible explanation might be that language assessment is not a topic that could be learned or acquired on the job, and there should be some extra driving forces for the teachers to have this knowledge. For experience to be helpful for a teacher's self-development, the teacher has to bring a lot of things to the classroom from her/his BA programme as the theoretical background. As the findings showed that the BA programme has no effect on LAK level

of the teachers, it is clear that the teachers start their jobs with insufficient knowledge in language assessment (Hatipoğlu, 2017). Moreover, if there are already going on practices in the institution in language assessment, because the teachers do not have adequate knowledge, they may just get used to the practices, even wrong practices. They may even do not realize that these are wrong practices, because, for this to be aware of, the teachers can be advised to follow literature to get theoretical knowledge, attend conferences specifically on language assessment or have role-models who are expert or more knowledgeable in language assessment.

The second variable is the BA programme being graduated from, namely, whether the teachers graduated from ELT programmes or not. The results indicated that the teachers who graduated from ELT departments and the teachers who were the graduates of non-ELT departments are not different in terms of their language assessment knowledge. In other words, whether teachers with ELT background or not does not play a role. At a first glance, both groups might be perceived as different in terms of their language assessment knowledge, because ELT programmes are specifically designed for foreign language education. However, there is not a difference between both groups. The reason for the similarities of both groups could be that language assessment is not given a priority in ELT programmes and covered in one course at the fourth grade; thus, the graduates of ELT and non-ELT are not different with respect to their language assessment knowledge.

With respect to educational background, there are no significant differences among the teachers having BA degree, MA degree and PhD degree. That is, even though the teachers have various educational background, their level of language assessment remains the same and does not change. This finding is in line with Tao's (2014) study, showing that there is not a statistically significant difference between the teachers whose educational background is BA and the ones with MA. As previously mentioned, the BA programmes were insufficient in terms of exposure in language assessment knowledge, and this finding might underline the situation that this insufficieny may not be solely the problem of pre-service education, but also might be the problem of MA and PhD programmes. Even if there is one course related to language assessment, compulsory or elective, in post-graduate level, all the topics have to be covered in one course in one academic term period, which might be short for this broad topic. These possible reasons may have led to this finding.

The other variable is workplace, whether the participant teachers work at a state or private university. The findings revealed that there is not a significant difference between the teachers working at a state and private university. Private universities tend to have more training and professional development programmes. However, the contents of these trainings and professional development programmes are crucial issues, because though there may be density of trainings and programmes, they may cover other issues related to language apart from language testing and assessment, or they may not cover language assessment in relation to the skills. Furthermore, the presence of the professional development programmes may not guarantee the increased knowledge in all fields of language including language testing and assessment, which is also one of the findings of the current study regarding the relationship between the existence of training and professional development programmes and language assessment knowledge.

The present study also investigated whether having a standalone course in pre-service education has an effect on language assessment knowledge of the participant teachers. However, the findings revealed that there is not a significant difference between the teachers who had a separate testing and assessment course in pre-service education and who did not. That is, the presence of that course in pre-service does not make the teachers more knowledgeable in terms of language assessment knowledge. The findings of this study could be supported by the findings of Tsagari (2008) and Tao (2014). In both studies, it was stated that the participants had inadequate assessment training in pre-service education. In Turkish context, this finding is in line with Köksal (2004)'s study which stated that pre-service education is insufficient concerning identifying characteristics of a good test, how to design, administer and score language tests. Additionally, the finding of the current study is in parallel with Mede and Atay (2017)'s study that aimed to explore assessment literacy levels of English teachers at foundation universities. It was revealed that pre-service education was found to be insufficient, and the teachers were in need of classroom-focused LTA domain, purposes of language assessment, and receptive, productive and integrated skill, because they had little training with respect to these topics in their pre-service education. Moreover, the participants found the content of the course too abstract, and they needed more opportunities for practice (Mede & Atay, 2017). This finding is also in parallel with Hatipoğlu (2015; 2017)'s studies in which the learners at a state university in Turkey uttered that testing and assessment course in pre-service education is not sufficient, and additions were

needed in that course. It is clear that although there is a specifically designed course in pre-service education for testing and assessment, this course is not enough and does not lead to increased language assessment knowledge of the teachers. As Hatipoğlu (2015) stated, one course in language testing and assessment in pre-service education resulted in lack of basic training of learners in language assessment. The insufficiency of this course in pre-service education may result from several probable causes. One might be related to the competency of the teacher educators giving those courses in pre-service education. These teacher educators should be equipped with a lot of knowledge related to language assessment. Stiggins (1999), Hatipoğlu (2015) and Jeong (2013) stated that the teacher educators who are responsible for this language assessment course at university should have a solid background in language assessment. The second one might be the arbitrariness of the content of these courses. There is not a framework for the syllabus design for these courses, and the teacher educators giving those courses decide on the content of these courses (Hatipoğlu, 2015). The third one is even though there is a specifically designed course, the presence of this course may not be enough to cover all the information related to assessing each language skill comprehensively in just one academic term period. The learners may not have sufficient time to become familiar with all the issues related to assessing language skills, and they may also not have time to make practice such as going through ready-made exams and deciding on the appropriacy of the tasks, or designing tasks. As they are not involved in these tasks, it is more likely that the presence of that separate course in pre-service education may not be very efficient for the teachers. The importance of practice was also stated by Jin (2010). The last reason might be related with the perspectives of teacher candidates. While taking this course, prospective teachers may think that the information covered in this course related to language assessment will not be useful for them. As they are not required to prepare or administer exams at that time, that is, it is not necessary for them to resort to their language assessment knowledge at that time, they may not fully comprehend the necessity and usefulness of language assessment knowledge, finally they may not give enough importance to that course.

The last variable which does not have an effect on the LAK level of the participant teachers is attending trainings on testing and assessment. In other words, attending trainings on testing and assessment does not necessarily mean that there will be an increase in LAK level of the teachers. This result is in contrast with Stiggins (2010) who

stated that the reason why teachers are assessment illiterate is lack of professional development programmes. Moreover, Mede and Atay (2017) stated that there is lack of training, and most of the participants in their study expressed that they had no or little training in skills. However, training in language assessment does not lead to increased knowledge in language assessment. McNamara and Roever (2006) drew attention to the weaknesses of trainings by stating that the trainings should go beyond applied psychometrics, and should have a comprehensive and to the point content. Malone (2008) also stated that training is not enough itself. Trainings should "include the necessary content for language instructors to apply what they have learned in the classroom and understand the available resources to supplement their formal training when they enter the classroom (p. 235). Attending trainings is not sufficient for a language teacher. As stated in Malone's sentences, trainings should go hand in hand with other efforts of language teachers. This inefficiency of the trainings or programmes on the language assessment knowledge of the teachers may rely on the fact that there are not many trainings or professional development programmes on language assessment, especially there is not a conference solely focusing on language assessment with respect to assessing four skills in Turkey. Another reason could be the sustainability of these programmes. Half of the participants in Mede and Atay's (2017) study stated that they had training in language assessment, but they were short and one-shot training. Hence, sustainability of the programmes may also play a role in increasing language assessment knowledge of teachers.

Being a testing office member is the only variable that makes a difference, and has an influence on language assessment knowledge of the participants. The results showed that there is a significant difference between the testing office members and the ones who are not in testing office. This finding was another focus of the qualitative data, and for the fourth question of the open-ended protocol, the participants were asked to comment on the significant impact of being a testing office member on language assessment knowledge of teachers. The respondents expressed that when teachers are more involved in assessment-related activities, they learn more. According to them, as testing office members have to be involved in assessment-related activities in testing office, they naturally learn more. On this involvement issue, two of the participants stated the following sentences.

"When you are in a testing office, you feel the pressure and need that you should be better and you should improve yourself in language testing and assessment. Now, as a part of this office, you are responsible for designing tests, writing items, etc."

"As testing office members, these teachers should know everything related to the question types, instructions, how to score the items in the tests they designed. They should have all this information because when a colleague or learner asks the logic behind them, they are expected to give an answer to the questions."

Another possible reason mentioned by the participants on testing office members' being more knowledgeable was related to the members' attending trainings and conducting research. The following statements expressed by one of the participants illustrate the issue well.

"As testing office members feel the need to be better in language testing and assessment, they tend to attend to conferences or conduct research, and their institution mostly encourages them for this professional development effort."

The opinions provided by the participants for the fourth question focusing on testing office members' being more knowledgeable clearly underlined the importance of being involved in testing and assessment practices for professional development in this area. It was mainly believed that it was the practices that made the difference between testing office members and non-members. Based on this, it can be concluded that whether being a testing office member or not, teachers working in higher education context should be encouraged to be involved in testing and assessment practices to reach a certain degree of competency in this domain.

5.5. Perceived Self-competency and Actual Language Assessment Knowledge Level

The fifth research question aimed to look into whether the teachers' LAK level changes according to their perceived self-competency in assessing each language skill. The results showed that there is not a significant difference among the participants who perceived themselves as very competent, competent, and not very competent in terms of their LAK level in assessing reading, listening, writing and speaking. Furthermore, the findings indicated that the majority of the participants perceived themselves competent or very competent. On the other hand, the ones who thought that they were not very competent in assessing each skill had the highest mean score among all. It can be concluded that the participant teachers' perceived self-competency in assessing these four skills is far from their actual LAK level. With respect to perceived self-competency, the

finding of this study shows parallelism with Öz and Atay's (2014) study in which 12 Turkish EFL teachers' beliefs concerning their in-class language assessment were investigated. The participant teachers reported that they were familiar with the concepts related to language assessment such as the features of a good test. In the same vein, in Jannati's (2015) study, the teachers stated that they had enough knowledge about the concepts and terminology related to language testing and assessment. The results of these studies are in parallel with the current study, and in all these studies, the teachers perceived themselves competent.

In line with the studies mentioned above, this finding shows the imbalance between the teachers' perceived self-competency and their actual LAK level. In the qualitative phase of the current study, the participant teachers were asked for the possible reasons of this inconsistency, and they stated that the teachers were not aware of their assessment knowledge levels. On this issue, one of the teachers uttered that:

> *"Teachers may think that what they experienced or learnt years ago was correct; so, they even do not feel the need to question their language assessment knowledge."*

One of the participants mentioned this unawareness by saying that:

> *"Going through the exam questions in the class with the students, and giving them the true answers and making them explanations on how to answer the questions do not mean that teachers are knowledgeable in assessment-related activities."*

The next reason of this mismatch could be the teachers' resistance to accept their incompetency. One participant touched upon this issue by saying that:

> *"I know that I have many weaknesses in assessment, but, most of the teachers do not want to accept this, and they say that it is not my favourite research area, or I am not a testing office member. However, each and every language teacher should have certain degree of language assessment knowledge. Moreover, most teachers do not have the willingness for self-reflection, and here is the result."*

The last reason expressed by the participants was the teachers' being unaware of the importance of language testing and assessment. One of the participants stated that:

> *"These assessment-related activities are thought to be the duties of testing office members. Thus, they may not find it necessary to learn the things related to language testing and assessment."*

The participants' answers to this question revealed that teachers, in general, are not aware of their weaknesses in language testing and assessment. This finding is also important in the sense that the first point to start with should be to increase the awareness

among teachers in higher education context regarding their level and weaknesses in language testing and assessment. Along with raising the teachers' awareness, resistance of the teachers should be dealt with, and they should be encouraged to be more open to assessment-related activities. They also should be informed about how crucial LTA is as a part of their profession.

5.6. Teachers' Needs on Language Assessment

The last research question of the study aimed to find out the needs of EFL teachers' on language testing and assessment, and what kind of a module they wanted to have in this area of expertise. Firstly, they were asked about their needs, and they stated that they definitely needed trainings and workshops for all skills, which are in line with the findings of Popham (2009) and Fulcher (2012). On this issue, one of the teachers expressed that:

> *"The important thing is that testing office members should be the focus of these trainings, and all the teachers in the institutions should not be included in these trainings. As a first step, testing office members should gather and decide on the topics on which they want to be trained, then based on this list, they should be given trainings. After the training of the testing office is over and testing office members become more competent in language testing and assessment in all skills, these trainings can also be given to all teachers working in the institution."*

Another point touched upon by the participants regarding their needs in language assessment was how to overcome subjectivity in productive skills. Popham (2009) also found out that the teachers had difficulty in scoring because of the subjective nature of certain skills and tasks. Moreover, the participants in Hasselgreen, Carlsen and Helness (2004) called for a training in productive skills. The sentences provided by one of the participants clearly expressed this need.

> *"Especially, in assessing reading and writing, I want to learn how to develop clear and to the point rubrics that decrease the subjectivity of scoring in those skills."*

The participants also stated that they wanted to learn how to construct tests regarding each skill in these trainings. Similarly, Hasselgreen, Carlsen and Helness (2004), Popham (2009), Wu (2014) and Mede and Atay (2017) concluded that the teachers need training regarding how to prepare tests. The participant teachers in the current study also wanted to learn how to analyze reliability and validity of tests, which was mentioned in Hasselgreen, Carlsen and Helness (2004), Wu (2004) and Popham's

(2009) studies, as well. Regarding these points, one teacher expressed the following sentences:

> *"We should be taught how to construct tests and tasks by the professionals. Knowing something and doing it correctly are different things; thus, I want to make a practice with the professionals based on the specific examples. Now, the numbers mean nothing to me, unfortunately. I want to analyze the reliability and validity of the tests we designed in our institution."*

The last open-ended question focused on the elements of a potential training module on LTA in the eyes of the participant teachers. When they were asked what kind of a training module they would like to have, they stated that in their ideal training module, the professional practitioners should give this training. The following sentences provide a good explanation for this.

> *"These trainings should be given by professionals who are involved in assessment practices regarding each skill. The problematic parts in which we cannot come to an agreement with our colleagues could be asked to the professionals and they should have the necessary knowledge and confidence to answer our questions."*

The participants also stated that this training module should be long-lasting and sustainable to create a significant impact. This finding shows parallelism with the results of Herrera and Macias (2015) which also revealed that ongoing training is a must to keep up with the recent innovations in LAL. One of the teachers mentioned this point by saying that:

> *"The trainings should be more beneficial when they are long-lasting and sustainable. Because, it is not very easy to learn new things or to adapt to new information. So, with the help of the recurrent trainings, teachers firstly become more aware of their practices, and start to apply what they have learned in those trainings."*

Another important element mentioned by the participants was hands-on practices. It was emphasized that such practices would definitely contribute to the development of the participants receiving a training module. Similarly, Lam (2015) also stated that training should include practices, and combine both theory and practice. Regarding this, one teacher expressed the following sentences:

> *"Teachers could work in small groups in these trainings, and when they have problems during practice, they can ask their questions to the professionals. Besides, with the help of the group work, teachers may have the opportunity to learn from each other."*

The last point expressed in open-ended questions was about the institutional elements. It was expressed that taking institutional factors into consideration during these trainings was quite important. Based on this, one respondent stated that:

> *"Not all the information in the trainings is applicable. Thus, the trainings should be context-specific, and train us by taking our institutional factors into consideration. Thanks to this, we could convert all this theory into practice."*

In general, the findings derived from this research question are in line with the studies conducted as needs analysis on language testing and assessment. It is seen that EFL teachers working in higher education context need training in language testing and assessment including all skills. Designing tasks and tests to assess all language skills, evaluating especially productive skills without being subjective, analyzing the validity and reliability of tests were some major elements regarding their needs. As for a potential training module, it was found that the participants did not favour theoretical training given by academicians. Instead, they expressed that they would prefer professional practitioners as trainers. Finally, sustainability, hands-on practices and certain institutional elements were also found as important elements to be included in a training module on LTA.

6. CONCLUSION

6.1. Summary of the Study

Assessment has a great role in teaching and learning process, and they are the two sides of a coin. They inform each other, and in turn, affect each other (Malone, 2013). Good assessment practices are important and necessary for the betterment of teaching and learning process. Teachers have great duties in assessment-related activities, because all these activities are carried out by teachers. As the burden on teachers' shoulders is big, they are expected to have a certain degree of assessment knowledge so that they can carry out assessment-related activities efficiently and can make use of their assessment knowledge in their practice. However, the problem is that teachers are in the center of all these assessment-related activities and they are responsible for assessing learners, but whether they are competent enough to carry out all these activities is open to discussion. There was an urgent need to develop and validate a scale measuring language assessment knowledge of teachers in ELT, because there is not such a measurement tool to the best knowledge of the researcher. Based on this need, the current study first aimed to develop and validate Language Assessment Knowledge Scale (LAKS) with 60 items and four skills that are reading, listening, writing and speaking, and investigated the psychometric properties of LAKS. The statistical analyses displayed a perfect model-data fit, and the reliability of each skill and LAKS in general was satisfactory. By means of this validated measurement tool, general and skill-based language assessment knowledge level of EFL teachers in Turkish higher education context were investigated, and it was seen that overall mean score out of 60 items was 25, meaning less than half of the items were answered correctly by the participants. When the skills were analysed in detail, it was clear that the highest mean score belonged to assessing reading, 7,055, and the lowest mean score belonged to assessing listening, 4,752. As the mean scores revealed, the participant teachers (N=542) were not knowledgeable enough in language assessment. The relationship among the teachers' skill-based assessment knowledge was examined, and it was observed that all skills were highly and positively correlated with LAK in general, and all skills were highly and positively correlated with each other as well. When the effects of demographic features on LAK level of the teachers were investigated, it was seen that the only significant difference was found among the participants in terms of being a testing office member or not. Testing office members were found to have higher mean scores. The rest of the demographic features, which are years of experience,

educational background, the BA programme being graduated, workplace, having a separate testing course in BA, and attending trainings on testing and assessment, were found to have no effect on LAK level of the participant teachers. Another scope of this study was to investigate the participants' perceived self-competency levels. It was revealed that their LAK level did not change according to their perceived self-competency level. In other words, there is not a statistically significant difference among the groups who perceive themselves very competent, competent and not very competent in terms of their language assessment knowledge.

What has been discussed till now is related to the findings of the quantitative data. As this study has a mixed design, qualitative data were also collected from 11 language teachers. The participants were asked to comment on the findings of the scale, and their needs related to language testing and assessment. As for the findings of this study, the respondents expressed that there was lack of knowledge in language testing and assessment, and they thought that why teachers were not knowledgeable enough in language assessment was due to insufficiency of pre-service and in-service education. The stated reasons for the insufficiency of pre-service education were limited exposure in the curriculum, teacher educators' insufficient knowledge, and non-ELT graduates. For them, insufficiency of in-service education resulted from insufficient professional development activities, lack of motivation of teachers, and lack of sources in language testing and assessment. The respondents were also asked to comment on the result that assessing reading got the highest mean score whereas listening got the lowest mean score. They stated that why assessing reading got the highest mean score was because of its priority in the curriculum, more concrete outcomes in reading, teachers' being more experienced with reading, and having more sources. Assessing listening was found to be challenging by the respondents because of practicality issues, the fact that teaching listening was not given enough importance, and insufficient experience in assessing listening. For the significant difference between testing office members and nontesting members, the participants uttered that when teachers were involved in assessment-related activities, they were more likely to learn more. They also commented on the perceived self-competency of the teachers and their actual language assessment knowledge level. The stated reasons were that the teachers were not aware of their assessment knowledge level and the importance of language testing and assessment, and they had a resistance to face with their incompetency. Finally, the last two research questions focused on the

opinions of the participants regarding the findings of the current study and their needs related to language testing and assessment. The findings indicated that they needed workshop and training regarding four skills, they wanted to learn how to overcome subjectivity in productive skills, how to construct tests and analyse their reliability and validity. When they were asked what kind of a training module they desired, they stated that the trainings should be given by professional LTA practitioners, should be loaded with practices, should be long-lasting and sustainable and institutional factors should be taken into consideration in these training modules.

6.2. Limitations of the Study

This study aimed to yield results regarding EFL teachers' general and skill-based language assessment knowledge levels. Even though many steps and procedures were followed systematically in this study, there are still certain limitations. One is the number of the participants answering open-ended questions. The open-ended questions were sent to 20 teachers; unfortunately, 11 of them replied the questions. It would have been better if there had been more teachers answering these questions. Also, it would have been better if interviews had been held to obtain more detailed data from the participants. Second one is the context in which the study was conducted. As the setting is limited to the preparatory programmes of the universities in Turkey, the results reflect the language assessment knowledge level of EFL teachers in higher education setting in Turkey. The last limitation is about some of the items in the scale with low factor loadings. It would have been better if all the items in the scale had had higher factor loadings, but these items were crucial for the content validity of the skills in the scale and for this reason, they were decided to be kept in.

6.3. Implications and Suggestions for Further Research

To the best knowledge of the researcher, this study is the first to develop and validate a measurement tool which is specifically designed to investigate general and skill-based language assessment knowledge level of EFL teachers. In the scope of this study, LAKS was developed and validated, and via LAKS, general and skill-based language assessment knowledge level of EFL teachers were revealed. Apart from the development of LAKS and its findings, the opinions of the participants based on the

findings of the scale and the needs of EFL teachers were also displayed. Based on all these findings, this study comes up with certain implications:

- As is clear from the findings of the current study, pre-service education has some limitations in terms of language testing and assessment, and pre-service teachers are not equipped with necessary knowledge in pre-service education related to language assessment. Thus, the content of the course in pre-service education might be considered to be revised. Moreover, one course cannot be sufficient for such a comprehensive topic to be covered in just one academic term. There should be more than one course related to language assessment, and more practical hands-on practice can be incorporated into these courses in pre-service education.
- Considering the relatively low level of language assessment knowledge of EFL teachers based on the findings, many efforts are needed to increase the language assessment knowledge level of EFL teachers, and to make them be more aware of how to make use of assessment-related activities more efficiently. Trainings and professional development programmes could be designed based on both theory and practice related to their needs, and language teachers could be supported and encouraged to attend the conferences and professional development programmes on language testing and assessment.
- A training module could be designed which is solely based on language testing and assessment regarding four skills. In this training module, teachers are given education regarding each skill in language assessment. In these training programmes, teachers could be provided with basic, practical and to the point information related to each skill, and they can work on real exams and could be asked to make comments on ready-made exams. Thus, they can have the chance to combine theory and practice, and the training becomes more meaningful. By identifying the weaknesses and strengths of ready-made exams, teachers could be involved in assessment-related activities, and whenever they have questions they can ask them to the trainers who are professional LTA practitioners. Furthermore, the attendance to these programmes for each language teacher should not be seen as something arbitrary, rather, it should be a must for each and every language teacher for their professional development.
- One of the problems related to pre-service education, as indicated by the participants of this study, was the incompetency of teacher educators in pre-

service education. To overcome this problem, it will be better if the teachers giving language testing and assesment courses can be specialized in this field and keep up with the literature concerning language assessment. Furthermore, attending the conferences or trainings on assessment could be a way of overcoming this problem. Moreover, as Higher Education Council does not provide a framework for these courses, the teachers from different universities can cooperate with each other to make these courses more fruitful for the learners.

- For the practical implication, the heads of the preparatory programs or the principles at schools could administer this scale to the language teachers working in their instutitions. Based on the findings derived from the scale, the heads or principals could detect the weaknesses and strengths of these participant teachers, and also could determine these teachers' needs regarding each skill. Based on their strengths, weaknesses, and needs of the teachers, professional development programmes could be determined and the teachers could be encouraged to attend to conferences, and regular meetings could be held to exchange information. Thus, language assessment knowledge of EFL teachers could be increased by being context-specific and taking institutional factors into consideration.

Above are the implications drawn from the findings of the current study. The results of this study have opened many doors for future studies. To start with, this study is restricted to the participants working at preparatory programmes of universities in Turkey. The same scale could be administered to the language teachers working at Ministry of Education and pre-service teachers in ELT departments. This measurement tool could also be used in other countries to indicate the language assessment level of language teachers. Besides, the reliability and validity studies on LAKS can be conducted in different settings for the development and adaptation of LAKS. In addition to these, some cultural, linguistic and context-specific elements could be added to the scale, and this could be carried out with teachers. Moreover, the problems teachers encounter in their assessment practices could be detected and they may be explored. Finally, in further studies, to what extent language teachers can make use of their language assessment knowledge in their practices could be searched. The tests they have designed or the tests they have scored could be investigated to see the similarities and differences between their language assessment knowledge and practices.

REFERENCES

Akbari, R. (2012). Validity in language testing. In C. Coombe, P. Davidson, B. O'Sullivan, and S. Stoynoff. (Eds.). *The Cambridge guide to second language assessment.* (pp. 30-36). Cambridge University Press, USA.

Alderson, J. C. (2000). *Assessing reading.* Cambridge: Cambridge University Press.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency.* London / New York: Continuum.

Alderson, J. C., and Wall, D. (1993). Does washback exist? *Applied Linguistics, 14* (2), 115-129.

Alkharusi, H. (2011). A logistic regression model predicting assessment literacy among in-service teachers. *Journal of Theory and Practice in Education, 7*(2), 280-291.

Al-Nouh, N. A., Taqi, H. A., and Abdul-Kareem, M. M. (2014). EFL primary school teachers' attitudes, knowledge and skills in alternative assessment. *International Educational Studies, 7* (5), 68-84.

Aryadoust, V. (2013). *Building a validity argument for a listening test of academic proficiency.* Newcastle: Cambridge Scholars Publishing.

Aslam, R. (1992). *Aspects of language teaching*. New Delhi: Northern Book Centre.

Aydın, B., Kızıltan, N., Öztürk, G., İpek, Ö. F., Yükselir, C., and Beceren, S. (2017). Optional English preparatory programs after HEC 2016 regulation: Opinions of asministrators on the current situation and problems. *Anadolu University Journal of Education Faculty*, *1* (2), 1-11.

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

Bachman, L. F., and Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Backlund, P., Brown, K., Gurry, J., and Jandt, F. (1980). Evaluating speaking and listening skill assessment instruments: Which one is best for you? *Language Arts, 57* (6), 621-627.

Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing, 13* (3), 257- 279.

Bailey, K. M. (1998). *Learning about Language Assessment: Dilemmas, Decisions, and Directions.* Boston: Heinle and Heinle.

Baker, B. A., and Riches, C. (2017). The development of EFL examinations in Haiti: Collaboration and language assessment literacy development. *Language Testing,* DOI: 10.1177/0265532217716732

Beverly, B.A., Tsushima, R., and Wang, S. (2014). Investigating Language Assessment Literacy: Collaboration between assessment specialists and Canadian university admissions officers. *CercleS, 4* (1), 137-157.

Black, P., and Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5* (1), 7-71.

Boyles, P. (2005). Assessment literacy. In M. H. Rosenbusch. (Ed.). *National assessment summit papers* (pp 18-24). Ames: IA Iowa State University.

Brady, L., and Kennedy, K. (2014). *Curriculum construction* (5th ed.). Frenchs Forest, NSW: Pearson Australia.

Bright, R. (2007). *Writing through the grades: Teaching writing in secondary schools.* Winnipeg, MB: Portage and Main Press.

Brown, H. D. (2003). *Language assessment: Principles and classroom practices.* Pearson Education.

Bütüner, S.Ö., Yiğit, N., ve Çimer, S.O. (2010). Ölçme değerlendirme okuryazarlığı envanterinin Türkçe'ye uyarlanması. *E- Journal of New World Sciences Academy, 5* (3), 792-809.

Buck. G. (2001). *Assessing listening.* Cambridge: Cambridge University Press.

Byrne, B. M. (2012). *Structural equation modeling with Mplus***.** Basic concepts, applications, and programming. Taylor and Francis Group, LLC.

Calderhead, J. (1996). Teachers: Beliefs and knowledge. In d. C. Berliner, and R. C. Calfee (Eds.), *Handbook of Educational Psychology* (pp. 709-725). New York: Macmillan.

Campbell, Y., Murphy, J. A., and Holt, J. K. (2002). *Psychometric analysis of an assessment literacy instrument: Applicability to preservice teachers*. Paper presented at the meeting of the Mid-Western Educational Research Association, Colombus, OH.

Carrell, R. L. (1998). Interactive text processing: Implications for ESL / second language reading classrooms. In P. L. Carrell, J. Devine and D. E. Eskey (Eds.). *Interactive approaches to second language reading* (pp. 239-259). Cambridge: Cambridge University Press.

Cash, M. M., and Schumm, J. S. (2006). Making sense of knowledge. Comprehending expository text. In J. S. Schumm (Ed.). *Reading assessment and instruction for all learners*, (pp. 262-296). New York, NY, US: Guilford Press.

Chan, Y. (2016). Elementary school EFL teachers' beliefs and practices of multiple assessments. *Reflections on English Language Teaching, 7* (1), 37-62.

Chapman, C. and King, R. (2009). *Differentiated instructional strategies for reading the content areas.* Thousand Oaks, CA: Corwin Press.

Clapham, C., and Corson, D. (Eds.) (1997). *Encylopedia of language and education, Volume 7: Language testing and assessment*. Dordrecht, Netherlands: Kluwer.

Cohen, A. D. (1994). *Assessing language ability in the classroom* (2nd ed.). New York: Heinle and Heinle.

Coombe, C. (2012). Second language assessment. In C. Coombe, P. Davidson, B. O'Sullivan, and S. Stoynoff. (Eds.). *The Cambridge guide to second language assessment*. (pp. 1-5). Cambridge University Press, USA.

Coombe, C., Davidson, P., O'Sullivan, B., and Stoynoff, S. (Eds.) (2012). *The Cambridge guide to second language assessment.* Cambridge, UK. Cambridge University Press.

Coombe, C., Troudi, S., and Al-Hamly, M. (2012). Foreign and second language teacher assessment literacy: Issues, challenges, and recommendations. In C. Coombe, P. Davidson, B. O'Sullivan, and S. Stoynoff. (Eds.). *The Cambridge guide to second language assessment*. (pp. 20-29). Cambridge University Press, USA.

Cromey, A., and Hanson, M. (2000). *An exploratory analysis of school-based student assessment systems*. North Central Regional Educational Laboratory Evaluation Reports. Retrieved from https://eric.ed.gov/?id=ED452221.

Creswell, J. W. (2012). Educational Research: Planning, conducting, and evaluating quantitative and qualitative research. Boston: Pearson Education.

Çokluk, Ö., Şekercioğlu, G. ve Büyüköztürk, Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları*. Pegem Akademi, Ankara.

Davidheiser, S.A. (2013). *Identifying areas for high school teacher development: A study of assessment literacy in the Central Bucks School District* (Unpublished PhD Dissertation). Drexel University, United States.

Davies, A. (1999). *Dictionary of Language Testing*. Cambridge, UK.

Davies, A. (2008). Textbook trends in teaching language testing. *Language Testing, 25* (3), 327-347.

DeLuca, C., Valiquette, A., Coombs, A., LaPointe-McEwan, D., and Luhanga, U. (2016). Teachers' approaches to classroom assessment: A large-scale survey. *Assessment in education: Principles, Policy and Practice*, 1-21.

DiRanna, K., Osmundson, E., Topps, J., Barakos, L., Gearhart, M.; Cerwin, K., Carnahan, D., and Strang, C. (2008). *Assessment-centered teaching: A Reflective Practice*. Thousand Oaks, CA: Corwin Press.

Douglas, D. (2010). *Understanding language testing.* London: Hodder.

Dörnyei, Z. (2007). Research methods in applied linguistics. New York: Oxford University Press.

Eskey, D. E. and Grabe, W. (1998). Interactive models for second language reading: Perspectives on instruction. In P. L. Carrell, J. Devine and D. E. Eskey (Eds.). *Interactive approaches to second language reading* (pp. 223-238). Cambridge: Cambridge University Press.

Falsgraf, C. (2005, April). Why a national assessment summit? New visions in action. National Assessment Summit. Meeting conducted in Alexanderia, Va. Retrieved from: http://www. nflrc.iastate.edu /nva/word documents/ assessment_2005/ pdf /nsap_introduction.pdf

Farhady, H. (2012). Principles of language assesment. In C. Coombe, P. Davidson, B. O'Sullivan, and S. Stoynoff. (Eds.). *The Cambridge guide to second language assessment*. (pp. 37-46). Cambridge University Press, USA.

Farrell, T. S. C. (2008). *Teaching Reading to English Language Learners: A reflective guide*. Thousand Oaks, CA: Corwin Press.

Flowerdew, J. (Ed.) (1994). *Academic listening: Research perspectives.* Cambridge: Cambridge University Press.

Flowerdew, J., and Miller, L. (2005). *Second language listening: Theory and practice.* New York: Cambridge University Press.

Flowerdew, J., and Miller, L. (2012). Assessing listening. In C. Coombe, P. Davidson, B. O'Sullivan, and S. Stoynoff. (Eds.). *The Cambridge guide to second language assessment*. (pp. 225-233). Cambridge University Press, USA.

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

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9* (2), 113-132.

Fulcher, G., and Davidson, F. (2007). *Language testing and assessment: An advanced resource book.* London: Routledge.

Fulcher, G., and Davidson, F. (2012). *Routledge handbook of language testing.* London and New York: Routledge.

Gelbal, S. (2013). *Ölçme ve Değerlendirme*. Anadolu Üniversitesi Açıköğretim Fakültesi Yayınları.

Green, A. (2014). *Exploring language assessment and testing: Language in action.* New York, NY:Routledge.

Gronlund, N. E. (1998). *Assessment of student achievement* (6th edition). Boston: Allyn and Bacon.

Gruba, P., and Corbel, C. (1997). Computer-based testing. In C. Clapham and D. Corson (Eds.), *Encylopedia of Language and Education: Volume 7: Language Testing and Assessment*, (pp. 141-149). Kluwer Academic Publishers, The Netherlands.

Hamp-Lyons, L. (1996). Applying ethical standards to portfolio assessment of writing in English as a second language. In M. Milanovic and N. Saville (Eds.), *Perfromance testing, cognition and assessment*, (pp. 151-162). Cambridge: Cambridge University Press.

Hamp-Lyons, L. (2006). Feedback in portfolio-based writing courses. In K. Hyland and F. Hyland (Eds.), *Feedback in Second Language Writing Contexts and Issues,* (pp. 140-161). London: Cambridge University Press.

Harris, D. P. (1969). *Testing English as a second language.* New York: McGraw-Hill.

Hasselgreen, A., Carlsen, C., and Helness, H. (2004). *European survey of language and assessment needs. Part one: General finding.* Retrieved April 22, 2018, from www.ealta eu.org/documents/resources/survey-report-pt1.pdf.

Hatipoğlu, Ç. (2015). 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. (2017). History of English language teacher training and English language testing and evaluation (ELTE) education in Turkey. In Y. Bayyurt and N. Sifakis (Eds). *English Language Education Policies and Practices in the Mediterranean Countries and Beyond* (pp. 227-257). Frankfurt: Peter Lang.

Haught J., and Crusan D. (2016). Filling the Gaps: L2 Grammar and Assessment Preparation for ELA Teachers. In L. de Oliveira, M. Shoffner (eds). *Teaching English Language Arts to English Language Learners* (pp. 171-192). Palgrave Macmillan, London

Heaton, J. B. (1990). *Writing English language tests.* (2nd ed.). Cambridge: Cambridge University Press.

Henning, G. (1987). *A guide to language testing, development, evaluation and research.* Cambridge, MA: Newbury House.

Herrera, L. and Macias, D. (2015). A call for language assessment literacy in the education and development of teachers of English as a foreign language. *Colomb. Appl. Linguist. J., 17*(2), 302-312.

Hu, L., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, *Structural Equation Modeling: A Multidisciplinary Journal, 6* (1), 1-55.

Hubley, N. N. (2012). Assessing reading. In C. Coombe, P. Davidson, B. O'Sullivan, and S. Stoynoff. (Eds.). *The Cambridge guide to second language assessment.* (pp. 211-217). Cambridge University Press, USA.

Hughes, A. (1989). *Testing for language teachers.* Cambridge: Cambridge University Press.

Huot, B. (1996). Toward a new theory of writing assessment. *College Composition and Communication, 47*, 549-566.

Impara, J. C. , Plake, B. S., and Fager, J. J. (1993). Teachers' assessment background and attitudes toward testing. *Theory into Practice, 32* (2), 113-117.

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing, 25* (3), 385-402.

Inbar-Lourie, O. (2013). Guest editorial to the special issue on language assessment literacy. *Language Testing, 30* (3), 301-307.

Inbar-Lourie, O. (2017). Language assessment literacy. In E. Shohamy, I. G. Or, and S. May (Eds), *Language Testing and Assessment* (pp. 1-14). Springer International Publishing.

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 professional preparation of foreign language teachers in China. *Language Testing, 27*(4), 555-584.

Karaman, P., ve Şahin, Ç. (2014). Öğretmen adaylarının ölçme değerlendirme okuryazarlıklarının belirlenmesi. *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi, 15* (2), 175-189.

Katz, A. (2012). Linking assesment with instructional aims and learning. *The Cambridge guide to second language assessment.* In C. Coombe, P. Davidson, B. O'Sullivan, and S. Stoynoff. (Eds.). *The Cambridge guide to second language assessment.* (pp. 66-73). Cambridge University Press, USA.

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), 1- 11.

Kremmel, B., and Harding, L. (forthcoming). Towards a comprehensive, empirical model of language assessment literacy.

Lam, R. (2015). Language assessment training in Hong Kong: Implications for language assessment literacy. *Language Testing, 32* (2), 169-197.

Lantolf, J. P. (2009). Dynamic assessment: The dialectic integration of instruction and assessment. *Language Teaching, 42*, 355-368.

Lantolf, J. P., and Thorne, S. (2006). *Sociocultural theory and the genesis of second language development.* Oxford: Oxford University Press.

Lee, I. (2017). *Classroom writing assessment and feedback in L2 school contexts.* Singapore: Springer.

Leung, C. (2014). Clasroom-based assessment issues for language teacher education. In A. J. Kunnan (Ed.), *The Companion to Language Assessment,* (pp. 1510-1519). Chichester, UK: Wiley Blackwell.

Lewkowicz, J. A. (1997). The integrated testing of a second language. In C. Clapham and D. Corson (Eds.), *Encylopedia of Language and Education: Volume 7: Language Testing and Assessment*, (pp. 75-85). Kluwer Academic Publishers, The Netherlands.

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

Mackey, A. and Gass, S. M. (2005). Second language research: Methodology and design. Mahwah, N. J.: Lawrance Erlbaum.

Madsen, H. S. (1983). *Techniques in testing.* New York and Oxford: Oxford University Press.

Malone, M. E. (2008). Training in language assessment. In E. Shohamy and N. Hornberger (Eds.), *Encylopedia of language and education, Vol. 7: Language testing and assessment* (2nd ed.), pp. 225-233). New York: Springer Science and Business Media.

Malone, M. E. (2013). The essentials of assessment literacy: Contrasts between testers and users. *Language Testing, 30* (3), 329-344.

Marzano, R. J. (2000). Transforming classroom grading. Alexandria, VA: Association for Supervision and Curriculum Development.

McInerney, D. M. (2014). *Educational Psychology: Constructing Learning*. Pearson: Australia

McKenna, M. C., and Stahl, K. A. D. (2015). *Assessment for reading instruction* (3rd ed.). New York: Guilford Press.

McMillan, J. H. (2000). Fundamental assessment principles for teachers and school administrators. *Practical Assessment, Reseacrh and Evaluation, 7*(8), 1-5.

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

McNamara, T. (1997). Performance testing. In C. Clapham and D. Corson (Eds.), *Encylopedia of Language and Education: Volume 7: Language Testing and Assessment,* (pp. 131-139). Kluwer Academic Publishers, The Netherlands.

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

Mead, N. A., and Rubin, D. L. (1985). *Assessing listening and speaking skills.* Retrieved from https://www. ericdigests.org/pre-923/speaking.htm.

Mede, E., and Atay, D. (2017). English Language Teachers' assessment literacy: The Turkish context. *Dil Dergisi, 168* (1), 1-5.

Mendoza, A. A., and Arandia, R. (2009). Language testing in Colombia: A call for more teacher education and teacher training in language assessment. *PROFILE, 11* (2), 55-70.

Mertler, A. C. (2003). Secondary teachers' assessment literacy: *Does classroom experience make a difference?. American Secondary Education, 33*(1), 49-64.

Mertler, C. A. (2009). Teachers' assessment knowledge and the perceptions of the impact of classroom assessment professional development. *Improving Schools, 12* (2), 101-113 DOI: 10.1177/1365480209105575

Mertler, C. A., and Campbell, C. (2005). *Measuring teachers' knowledge and application of classroom assessment concepts: Development of the assessment literacy inventory*. Paper presented at the annual meeting of the American Research Association, Montreal, Quebec, Canada. Retrieved from https://eric.ed.gov/?id=ED490355.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13* (3), 241-256.

Morley, J. (1972). *Improving aural comprehension.* Ann Arbor, MI: Unıversity of Michigan Press.

Nunan, D., and Miller, L. (1995). *New ways in teaching listening.* Alexandria, VA: TESOL.

Oller, J. W. (1979). *Language tests at school: A pragmatic approach.* London:Longman.

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

O'Sullivan, B. (2012) Assessing speaking. In C. Coombe, P. Davidson, B. O'Sullivan, and S. Stoynoff. (Eds.). *The Cambridge guide to second language assessment.* (pp. 234-246). Cambridge University Press, USA.

Osterlind, S. J. (1989). *Constructing test items.* Boston: Kluwer Academic Publishers.

Öz, H. (2014). Turkish teachers' practices of assessment for learning in the English as a foreign language classroom. *Journal of Language Teaching and Research, 5* (4), 775-785.

Öz, S., and Atay, D. (2017). Turkish EFL teachers' in-class language assessment literacy: perceptions and practices. *ELT Research Journal, 6* (1), 25-44.

Pearson, P. D., and Johnson, D. D. (1978). *Teaching reading comprehension.* New York, NJ: Holt, Rinehart and Winston.

Perfetti, C.A., and Adlof, S. M. (2012). Reading comprehension: A conceptual framework for word meaning to text meaning. In J. Sabatini, E. Albro, and T. O'Reilly (Eds.). *Measuring up: Advances in how we assess reading ability* (pp. 3-20). Lanham, MD: Rowman and Littlefield Education.

Pill, J., and Harding, L. (2013). Defining the language assessment literacy gap: Evidence from a parliamentary inquiry. *Language Testing, 30* (3), 381-402.

Plake, B. S. (1993). Teacher assessment literacy: Teachers' competencies in the educational assesment of students. *Mid-Western Educational Researcher, 6* (1), 21-27.

Popham, W. J. (2004). 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, J. W. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory Into Practice, 48*, 4-11.

Price, M., Rust, C., O'Donovan, B., Handley, K., and Bryant, R. (2012). *Assessment literacy: The foundation for improving student learning.* Oxford: The Oxford Centre for Staff and Learning Development.

Purpura, J. E. (2004). *Assessing grammar.* Cambridge: Cambridge University Press.

Purpura, J. E. (2016). Second and foreign language assessment. *The Modern Language Journal,* DOI: 10.1111/modl.12308

Rao, Z., and Li, X. (2017). Native and non-native teachers' perceptions of error gravity: The effects of cultural and educational factors. *The Asia-Pacific Education Researcher, 26* (1), 51-59.

Read, J. (2012). Assessing vocabulary. In C. Coombe, P. Davidson, B. O'Sullivan, and S. Stoynoff. (Eds.). *The Cambridge guide to second language assessment*. (pp. 257-264). New York: Cambridge University Press.

Rost, M. (1990). *Listening in language learning.* Harlow: Longman.

Rost, M. (2011). *Teaching and researching listening* (3rd edition). Harlow, UK: Pearson Education.

Ruth, L., and Murphy, S. (1988). *Designing writing tasks for the assessment of writing.* Norwood, NJ: Ablex.

Salend, S. J. (2009). *Classroom testing and assessment for ALL students: Beyond standardization.* Thousand Oaks, CA: Corwin.

Salim, B. (2001). *A companion to teaching of English.* Atlantic Publishers, New Delhi.

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.

Schmitt, N. (1993). Comparing native and non-native teachers' evaluation of error seriousness. *JALT Journal, 15* (2), 181-191.

Sellan, R. (2017). Developing assessment literacy in Singapore: How teachers broaden English language learning by expanding assessment constructs. *Papers in Language Testing and Assessment, 6* (1), 64-87.

Shaw, S. D., and Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge: Cambridge University Press.

Sheorey, R. (1986). Error perceptions of native speaking and non-native speaking teachers of ESL. *ELT Journal, 40* (4), 306-312.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational researcher, 29* (7), 4-14.

Shrock, S. A., and Coscarelli, W. C. (2007). Criterion-referenced test development: Technical and legal guidelines for corporate training. (3rd ed.). San Francisco, CA: Wiley.

Silva, T. (1990). Second language composition instruction: Developments, issues and directions in ESL. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 11-36). New York: Cambridge University Press.

Stanford, P., and Reeves, S. (2005). Assessment that drives instruction. *Teaching Exceptional Children, 37* (4), 18-22.

Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan, 72*, 534-539.

Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan, 77* (3), 238-245.

Stiggins, R. J. (1999). Evaluating classroom assessment training in teacher education programs. *Educational Measurement: Issues and Practice*, 18 (1), 23-27.

Stiggins, R. J. (2007). Conquering the formative assessment frontier. In J. H. McMillan (Ed.). *Formative classroom assessment: Theory into Practice* (pp. 8-28). New York, NY: Teachers College Press.

Stiggins, R. J. (2010). Essential formative assessment competencies for teachers and school leaders. In H. L. Andrade, G. J. Cizek (Eds.), *Handbook of formative assessment,* (pp. 233-250). New York, NY: Taylor and Francis.

Stoynoff, S., and Coombe, C. (2012). Professional development in language assessment. In C. Coombe, P. Davidson, B. O'Sullivan, and S. Stoynoff. (Eds.). *The Cambridge guide to second language assessment*. (pp. 122-130). New York: Cambridge University Press.

Tao, N. (2014). *Development and validation of classroom assessment literacy scales: English as a Foreign Language (EFL) teachers in a Cambodian Higher Education Setting.* Unpublished PhD dissertation, Victoria University, Australia.

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics, 29*, 21-36.

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing, 30*(3), 403-412.

The American Federation of Teachers, the National Council on Measurement in Education, and the National Education Association (1990). *The standards for teacher competence in the educational assessment of students.* Retrieved from http://buros.org/standards-teacher-competence-educational-assessment-students

Thomas, J., Allman, C., and Beech, M. (2004). Assessment for the diverse classroom: A handbook for teachers. Tallahassee, FL: Florida Department of Education, Bureau of Exceptional Education and Student Services. Retrieved from http: //www.fldoe,org/ese/pdf/assess_diverse,pdf.

Thompson, I. (1995). Assessment of second/foreign language listening comprehension. In d. J. Mendelson, and J. Rubin (Eds.). *A guide for the teaching of second language listening.* San Diego, CA: Dominie Press.

Tsagari, D. (2008). *Assessment literacy of EFL teachers in Greece: Current trends and future prospects.* PowerPoint presentation at the 5th Annual EALTA Conference, May 9-11, Athens, Greece.

Tsagari, D. and Vogt, K. (2017). Assessment Literacy of Foreign Language Teachers around Europe: Research, Challenges and Future Prospects. *Papers in Language Testing and Assessment*, 6 (1), 41-64.

Valette, R. (1977). *Modern language testing*. New York: Harcourt Brace.

Volante, L., and Fazio, X. (2007). Exploring teacher candidates' assessment literacy: Implications for teacher education reform and Professional development. *Canadian Journal of Education,* 30 (3), 749-770.

Wall, D. (2012). Washback. In G. Fulcher and D. Fulcher (Eds.). *Routledge handbook of language testing* (pp. 79-92). London and New York: Routledge.

Webb, N. L. (2002). Assessment literacy in a standards-based urban education setting. A paper presented at the American Educational Research Association Annual Meeting. New Orleans, Louisiana. Retrieved from http:// citeseerx.ist.psu.edu/ viewdoc/download? doi=10.1.1.573.676&rep=rep1&type=pdf

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

Weigle, S. C. (2012). Assessing writing. In C. Coombe, P. Davidson, B. O'Sullivan, and S. Stoynoff. (Eds.). *The Cambridge guide to second language assessment*. (pp. 236-246). New York: Cambridge University Press.

Weir, C. J. (1990). *Communicative language testing*. Prentice Hall. Indiana

Weir, C. J. (1993). *Understanding and developing language tests.* New York: Prentice Hall.

West, R. and Turner, L. H. (2009). *Understanding interpersonal communication: Making choices in changing times.* Boston, MA: Wadsworth Cengage Learning.

White, E. (2009). Are you assessment literate?. *OnCue Journal, 3* (1), 3-25.

Wu, J. R. W. (2014). Investigating Taiwanese teachers' language testing and assessment needs. *English Teaching and Learning, 38* (1), 1-27.

Xu, Y., and Liu, Y. (2009). Teacher Assessment Knowledge and Practice: A narrative inquiry of a Chinese college EFL teacher's experience. *TESOL Quarterly 43* (3), 493-513.

Xu, Y., and Brown, G.T.L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education*, *58*, 149-162.

Xu, Y., and 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.

Yıldırım, A., and Şimşek, H. (2016). *Sosyal bilimlerde nitel araştırma yöntemleri*. Ankara: Seçkin.

Yüce, Z. (2015). *Pre-service English language teachers' conceptions of assessment and assessment practices.* (Unpublished Master's Thesis), Çağ University, Turkey.

Zhang, Z., and Burry-Stock, J. A. (2003). Classroom assessment practices and teachers' self-perceived assessment skills. *Applied Measurement in Education, 16* (4), 323-342.

APPENDIX A

LANGUAGE ASSESSMENT KNOWLEDGE SCALE – LAKS

PART I: DEMOGRAPHIC INFORMATION

1. Gender
a) male **b)** female

2. Years of experience

a)1-5 years **b)** 6- 10 years **c)** 11- 15 years
d) 16- 20 years **e)** more than 21years

3. The BA programme you graduated from
a) English Language Teaching (ELT) **b)** non- ELT

4. Educational background
a) BA degree **b)** MA degree **c)** PhD degree

5. Where are you working at now?
a) a state university **b)** a private university

6. Have you ever been a member of a testing office?
a) yes **b)** no

7. Did you have a separate testing/assessment course in pre-service education?
a) yes **b)** no

8. Have you attended any professional development programmes/ courses/ training on language assessment?
a) yes **b)** no

9. How do you evaluate yourself as an assessor in the following areas/subskills?

a) reading (1) very competent (2) competent (3) not very competent (4) not competent

b) listening (1) very competent (2) competent (3) not very competent (4) not competent

c) writing (1) very competent (2) competent (3) not very competent (4) not competent

d) speaking (1) very competent (2) competent (3) not very competent (4) not competent

10. Please write your institutional e-mail address: ……..

PART II: LANGUAGE ASSESSMENT KNOWLEDGE SCALE

ITEMS	True	False	Don't Know
ASSESSING READING			
1. Asking learners to summarize the reading text is a way of assessing their reading skills.			
2. When asking several questions about a reading text, all the questions are independent of each other.			
3. Cloze test is used for assessing the main idea of the text.			
4. In a reading exam, using a text learners have encountered before is not a problem.			
5. One reading text is enough to be included in a reading exam.			
6. The language of the questions is simpler than the text itself.			
7. Errors of spelling are penalized while scoring.			
8. Taking vocabulary difficulty into consideration is necessary in assessing reading skills.			
9. Including not stated/doesn't say along with true/false items has advantages over true/false items.			
10. The more items a reading text is followed, the more reliable it becomes.			
11. Using the same words in the correct option as in the text is not a problem.			
12. Simplification of reading texts is avoided.			
13. Reading texts in a reading exam include various genres (essay, article, etc.).			
14. In top-down approach, assessment is on overall comprehension of the reading text.			
15. Using ungrammatical distractors in multiple choice questions in a reading exam is a problem.			
ASSESSING LISTENING			
16. Using reading texts for listening purposes poses a problem.			
17. Including redundancy (e.g. what I mean to say is that ….) in a listening text poses a problem.			
18. Any type of listening text is used for note-taking.			
19. Spelling errors are ignored in scoring the dictation.			
20. Errors of grammar or spelling are penalized while scoring.			
21. A listening cloze test is a way of selective listening.			
22. Phonemic discrimination tasks (e.g. minimal pairs such as sheep-ship) are examples of integrative testing.			
23. Scoring in note-taking is straightforward.			

24. In discrete-point testing, comprehension is at the literal/local level.			
25. Using dictation diagnostically in assessing listening skills does not pose a problem.			
26. Giving learners a transcript of the listening text is a valid way of assessing listening skills.			
27. Dictation is a kind of discrete-point testing.			
28. Inference questions based on intelligence are avoided in listening tests.			
29. Asking learners to listen to names or numbers is called intensive listening.			
30. In selective listening, learners are expected to look for certain information.			
ASSESSING WRITING			
31. Giving two options to learners and asking them to write about one ensure reliable and valid scoring.			
32. Analytic scoring is used to see the strengths and weaknesses of learners.			
33. The parts of a scoring scale and the scores in each part do not change for different levels of learners.			
34. When there is a disagreement between the scores of the two raters, they score the written work again.			
35. Learners are required to write about at least two tasks in the exam rather than one task.			
36. Giving restrictive prompts/guidelines to learners for the writing task is avoided.			
37. Giving learners an opinion and asking them to discuss it is a valid way of assessing their writing skills.			
38. Using visuals which guide learners for writing poses a problem.			
39. Holistic scoring is used to see whether the learner is proficient or not at the end of the term.			
40. Analytic scoring leads to greater reliability than holistic scoring in writing.			
41. In controlled writing, learners have the chance to convey new information.			
42. Classroom evaluation of learning in terms of writing is best served through analytic scoring rather than holistic scoring.			
43. Irrelevant ideas are ignored in the assessment of initial stages of a written work in process writing.			
44. Providing a reading text for writing is a way of assessing writing skills.			
45. Mechanical errors (e.g. spelling and punctuation) are dealt with in the assessment of later stages of a written work.			

ASSESSING SPEAKING			
46. When the interlocutor does not understand the learner, giving that feeling or saying it poses a problem.			
47. Giving learners one task is enough to assess speaking skills.			
48. Interlocutors' showing interest by verbal and non-verbal signals poses a problem.			
49. When it becomes apparent that the learner cannot reach the criterion level, the task is ended.			
50. Using holistic and analytic scales at the same time poses a problem.			
51. Reading aloud is a technique used to assess speaking skills.			
52. In interlocutor-learner interviews, the teacher has the chance to adapt the questions being asked.			
53. In interactive tasks, more than two learners pose a problem.			
54. The interlocutor gives the score when the learner is in the exam room.			
55. In a speaking exam, production and comprehension are assessed together.			
56. Asking learners to repeat a word, phrase or a sentence is a way of assessing speaking skills.			
57. Discussion among learners is a way of assessing speaking skills.			
58. A checklist is a means of scoring oral presentations in in-class assessment.			
59. When the focus is to assess discourse, role plays are used.			
60. In peer interaction, random matching is avoided.			

APPENDIX B

OPEN-ENDED QUESTIONS PROTOCOL

Dear Participant,

The answers you give to the following questions will be evaluated within the qualitative data of the doctoral thesis I have been pursuing in the Department of English Language Teaching at Anadolu University. We developed a scale called "Language Assessment Knowledge Scale" to collect the quantitative data of my dissertation. In total, 542 teachers working at the preparatory programmes at school of foreign languages in Turkey completed the scale. Some of the questions below are about the results obtained from the scale, and the others focus on your opinions regarding language testing and assessment in general.

Your answers will only be used for this study. It is important that you answer the questions as detailed as possible so that we can learn your ideas.

For your inquiries, please do not hesitate to write to elcinolmezerozturk@anadolu.edu.tr

We thank for your support and contribution to our study.

Inst. Elçin ÖLMEZER-ÖZTÜRK

Supervisor: Prof. Dr. Belgin AYDIN

QUESTIONS

Are you a testing-office member Yes ( ) No ( )

1. According to "**Language Knowledge Assessment Scale**" developed within the scope of this study, language assessment knowledge level of the teachers working at the schools of foreign languages was identified as 25 out of 60. How do you evaluate this situation? What might be the underlying reasons of this situation?
2. There are four sections in the scale, assessing reading, listening, writing and speaking, each consisting of 15 questions. In terms of assessing the skills, the highest knowledge level was found in assessing reading (7.05) whereas the lowest level was in assessing listening (4.75). The knowledge level in assessing other skills was found as 6.80 in speaking and 6.57 in writing. How do you evaluate this situation? What are the possible reasons of this?

3. In the study, whether language assessment knowledge of the teachers changed according to different demographic characteristics that are years of experience, educational background, the BA programme being graduated, working at a private or state university, having a testing course in BA, and attending trainings on testing and assessment was investigated, and it was seen that none of them had an influence on their knowledge. How do you evaluate this?

4. The only significant difference was found between the participants who worked as testing office members and who did not. How do you interpret this difference and the potential reasons of it?

5. The relationship between the participants' perceived self-competency and their actual knowledge level was searched, and it was seen that most of them perceived themselves as competent or very competent although their actual score was 25 out of 60. How do you evaluate this difference? What can be the potential reasons of it?

6. What do you think your needs are in terms of your knowledge in assessing each skill?

7. What kind of an in-service training module do you think will meet your needs?

APPENDIX C

ETİK KURUL İZNİ

Evrak Kayıt Tarihi: 14.04.2017 | Protokol No: 44830

Tarih: 24.04.2017

ANADOLU ÜNİVERSİTESİ
SOSYAL VE BEŞERİ BİLİMLER BİLİMSEL ARAŞTIRMA VE YAYIN ETİĞİ KURULU
KARAR BELGESİ

ÇALIŞMANIN TÜRÜ:	BAP Projesi-Doktora Tez Çalışması
KONU:	Eğitim Bilimleri
BAŞLIK:	Dilde Ölçme Değerlendirme Okuryazarlığı: Dilde Ölçme Değerlendirme Ölçeğinin Geliştirilmesi ve Türkiye'de Yüksek Öğrenim Kurumlarında Çalışan Dil Öğretmenlerinin Dilde Ölçme Değerlendirme Bilgisinin Araştırılması
~~PROJE~~/TEZ YÜRÜTÜCÜSÜ:	Doç. Dr. Belgin AYDIN
TEZ YAZARI:	Elçin GİLMEZER ÖZTÜRK
ALT KOMİSYON GÖRÜŞÜ:	-
KARAR:	Olumlu

Prof.Dr. Çağlar BAYRAK (Başkan-Eğitim Fak.)	
Prof.Dr. Volkan YÜZER (Başkan Yardımcısı-Açıköğretim Fak.)	Prof.Dr. Esra CEYHAN (Eğitim Fak.)
Prof.Dr. Müneyyer ÇAKI (Güzel Sanatlar Fak.)	Prof.Dr. M. Erkan ÜYÜMEZ (İkt. ve İdari Bil. Fak.)
Prof.Dr. Handan DEVECİ (Eğitim Fak.)	Prof.Dr. Emel ŞIKLAR (İkt. ve İdari Bil. Fak.)